Technical Report

NetApp High-Performance Computing Solution for Lustre: Solution Guide

Robert Lai, NetApp
August 2012 | TR-3997

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 Introduction

1.1 NetApp HPC Solution for Lustre Introduction

Overview

The NetApp® High-Performance Computing (HPC) Solution for Lustre provides high-capacity and high-performance E-Series storage platforms that enable the Lustre™ file system to support very large scalability and extremely high input/output (I/O) throughput in modeling and simulation environments. The scalable and highly reliable design provides the ability to meet current and future requirements for performance and growth.

Figure 1 illustrates the stages of typical Lustre workflows.

Figure 1) Typical Lustre workflows.

The NetApp HPC Solution for Lustre, based on the E-Series platform, is purpose-built for scalable, reliable, and high-performance computational requirements for extreme I/O performance and massive file system capacity. Government, university, research, and business organizations will find that the NetApp HPC Solution for Lustre meets the challenge of supporting tens of thousands of Lustre clients accessing hundreds of petabytes of storage with I/O throughput of thousands of gigabytes per second.

Architecture and Components

The NetApp HPC Solution for Lustre consists of E-Series storage and the Lustre global parallel file system for object storage targets (OSTs) and metadata targets (MDTs). NetApp Professional Services and SupportEdge are required solution components.

The NetApp HPC Solution for Lustre is composed of the components listed in Table 1.

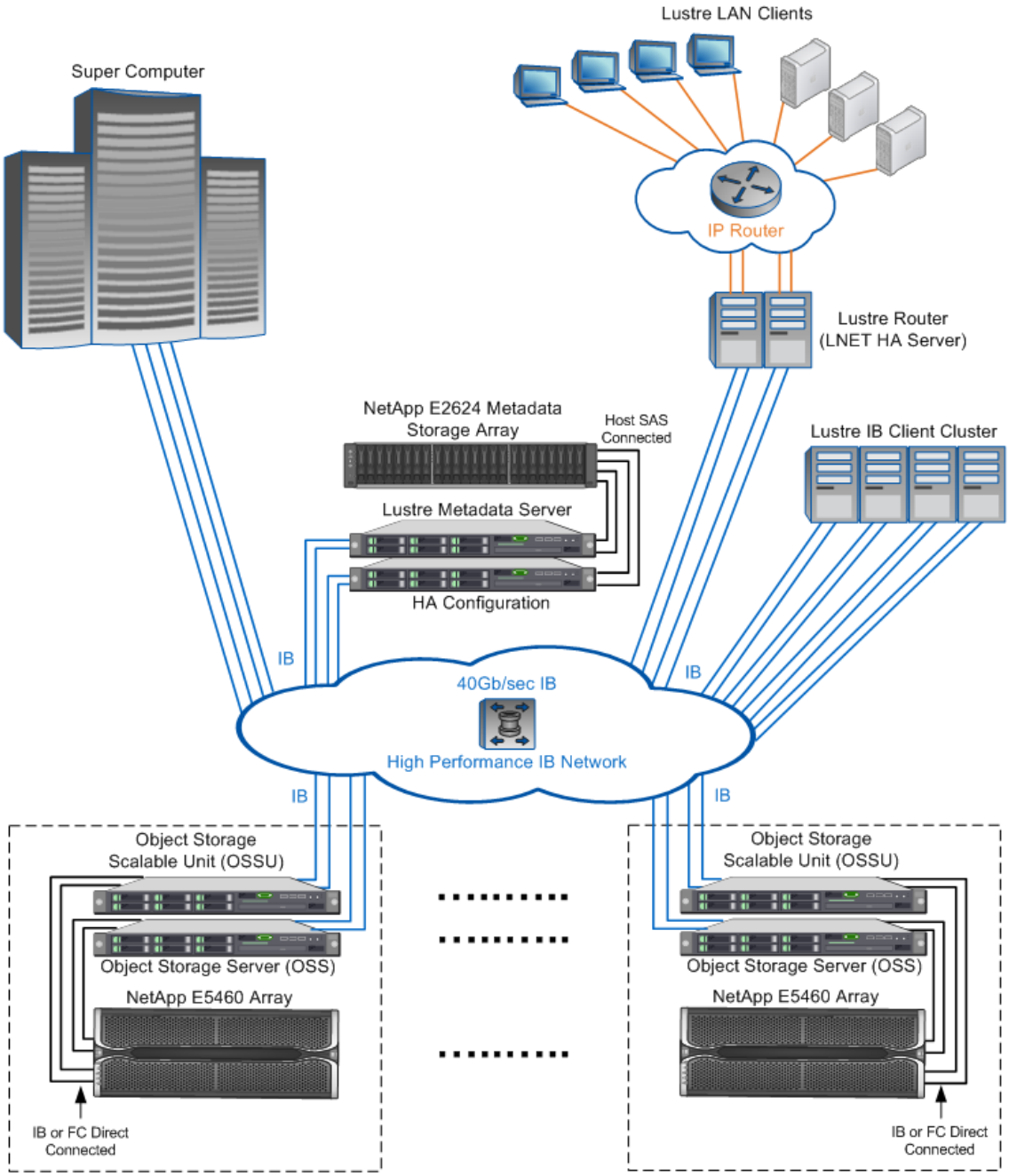
Table 1) NetApp HPC Solution for Lustre component list.

Component	Description	Hardware	Supplier
Object storage server (OSS)	The OSSs provide file I/O services to clients and manage data on the OSTs. OSSs are typically connected directly to E-Series-based OSTs by using Fibre Channel (FC) or InfiniBand (IB) protocols.	Linux® server (two for high-availability [HA] mode)	Integrator or customer
Object storage target (OST)	The OST is a Lustre logical unit number (LUN) residing on an E5460- or E5424-based storage device, providing Lustre object storage space.	NetApp E5460, E5424	NetApp
Object storage scalable unit (OSSU)	The OSSU is a scalable building block that composes an HA pair of OSSs and associated OSTs.	NetApp E5460/E5424 plus OSS	NetApp plus integrator or customer
Metadata server (MDS)	The MDS provides metadata services to clients. The MDS is connected to the metadata storage by using the SAS protocol.	Linux server (two for HA mode)	Integrator or customer
Metadata target (MDT)	The MDT is the storage for housing the file system metadata. Metadata storage is separate from the application/user storage. The MDT is typically connected to the MDS by using SAS.	NetApp E2624	NetApp
Management server (MGS)	The MGS manages the file system configuration.	Linux server	Integrator or customer
Lustre clients	These clients run applications that are network-attached to the Lustre client fabric and are the computational or I/O nodes for the Lustre application.	Linux servers	Integrator or customer
Lustre client fabric	IB, Ethernet (or other cluster fabric), and corresponding cabling to provide I/O access between Lustre clients and OSS servers.	IB network or Ethernet network (or other)	Integrator or customer
Second- and third-tier storage	Additional and optional methods for archiving data to disk or tape.	Near-line (NL) disk storage and/or tape library archive	Integrator or customer

For more information, refer to the datasheet for the NetApp E5400 Storage System.

Figure 2 provides a simplified overview of the NetApp HPC Solution for Lustre in a computational and visualization environment that uses an IB Lustre client fabric.

Figure 2) Typical Lustre architecture.



E-Series Architecture

The NetApp HPC Solution for Lustre consists of the E5460 and/or the E5424 E-Series storage systems. These storage systems feature dual E5400 RAID controllers in either the DE6600 4U 60-drive shelf or the DE5600 2U 24-drive shelf. Each shelf can be populated with NL-SAS, SAS, or solid-state drives (SSDs).

The E5460 and E5424 are fifth-generation storage arrays that include patented mechanical engineering and provide dense, scalable, and highly reliable bandwidth and capacity. The disk controller firmware supports an optimal mix of high-bandwidth, large-block streaming and small-block random I/O.

A base E5460 or E5424 may be expanded with the addition of one or multiple corresponding DE6600 or DE5600 expansion enclosures. The DE6600 and DE5600 are disk expansion enclosures or shelves that hold disks but no RAID controllers. These are cabled to the E5460 or E5424 and provide expansion storage behind the RAID controllers in the base unit. Thus, the NetApp HPC Solution for Lustre can be architected to independently scale capacity and bandwidth to best meet customer requirements.

The solution also uses an E2624 to store the Lustre file system metadata. The E2624 consists of a DE5600 shelf but uses the E2600 RAID controller instead of the E5400.

Table 2 provides guidelines on E-Series expansion options for the NetApp HPC Solution for Lustre.

Table 2) E-Series expansion guidelines for the NetApp HPC Solution for Lustre.

Category	E5460	E5424	E2624
Form factor	4U/60 drives	2U/24 drives	2U/24 drives
Maximum disk drives	360	192	24
Controller shelf	1	1	1
Maximum expansion shelves	5	7	0*
Total number of disk shelves	6	8	1

*The E2624 may be expanded with up to seven additional shelves, but the NetApp HPC Solution for Lustre is specified with the E2624 controller shelf only.

Lustre File System Architecture

The Lustre file system is a massively parallel file system that is capable of scaling to hundreds of petabytes, with I/O throughput of thousands of gigabytes per second while servicing thousands of clients. Its single file system namespace provides concurrent read/write access to multiple clients with a distributed lock manager to provide consistent file coherency between all clients.

The OSS provides file I/O services to clients by managing user data on OSTs. The OSTs are provided as LUNs from the E-Series storage system connected to the OSSs.

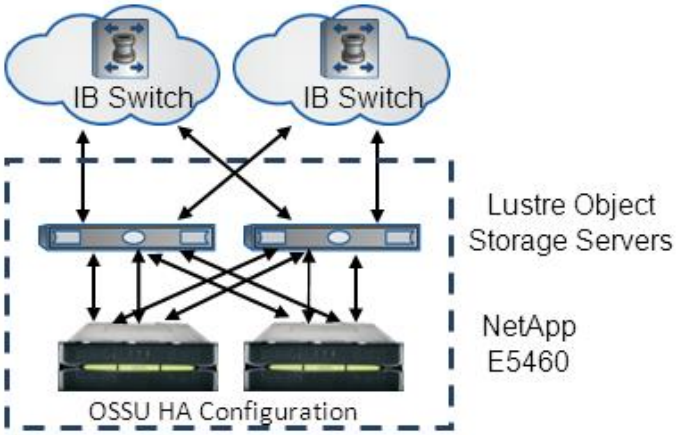
The NetApp OSSU combines two OSSs with their associated E-Series 5424/5460 systems (and corresponding OSTs) into a storage building block that is used to scale Lustre file systems for performance, capacity, and high availability.

The basic OSSU design corresponds to a pair of OSSs that are configured to a single E-Series storage system. This design provides the basic HA feature and guards against OSS failures that might reduce access to Lustre user data. The OSSU design prevents a single OSS failure from damaging user access to data. Each OSSU is a standalone unit and does not share storage connectivity or performance with any other OSSU.

An alternative basic OSSU design involves two OSSs configured to two E5424/E5460 storage systems, along with any associated expansion shelves. Each of the two E5424/E5460 storage systems splits its OST complement to provide OST ownership and control to each of the two OSSs (each OSS controls half the OSTs in each of the two E-Series 5424/5460 systems) as shown in Figure 3. This configuration enables the scale-out of Lustre OSSUs; the combination of OSSs to OSTs allows throughput, capacity, and OSS control to scale optimally.

The NetApp HPC Solution for Lustre configuration best practice refers to a pair of OSSs and their associated OSTs as an OSSU. The Lustre file system scales by adding OSSUs as required, achieving overall performance and capacity. Figure 3 shows the NetApp recommended HA OSSU architecture.

Figure 3) Example of an OSSU HA configuration.



These OSSUs are connected to the overall Lustre file system through connections from the OSSs to the Lustre client fabric network, greatly simplifying the Lustre scale-out and the overall design. Scale-out of the file system is achieved by presenting an incremental number of OSSUs on the client fabric through a single file system namespace.

The NetApp HPC Solution for Lustre characterizes each OSSU design for performance and capacity. The customer-specific Lustre file system is designed by using the appropriate number of OSSUs to achieve the overall performance and capacity criteria as scaled from the single OSSU characterized values.

The MDS manages Lustre metadata and stores the data in MDTs. This MDS provides the file layout mapping to Lustre clients for the collection of OSSUs in the file system. A second MDS provides high availability and is recommended for all configurations.

Conclusion

The NetApp HPC Solution for Lustre is designed and optimized for the most demanding computational and visualization processing workloads. This preconfigured, pretested solution is designed to support the high bandwidth required to process large volumes of data for large numbers of users. By enabling faster data processing, organizations are able to better support technical and business computing modeling and simulation:

- Big bandwidth support delivers up to 3.5GB/sec bandwidth in a single 4U rack unit.
- Modular design allows growth with minimal components, eliminating the need to overconfigure.
- High density supports up to 1.8PB in each industry-standard 40U rack.
- Cost-effective expansion allows scaling of bandwidth and capacity independently within the same container. Start small and expand with 2U or 4U increments as demand changes.

2 Solution Overview

2.1 NetApp HPC Solution for Lustre Sizing Considerations

Overview

Sizing is a critical component of architecting a NetApp HPC Solution for Lustre. It begins with the ability to meet specific goals of capacity and throughput. To properly size a Lustre file system, it is important to gather as much information and as many requirements as possible. These are categorized as follows:

- Infrastructure environment, user population, applications, and workflows
- Application data storage capacity required by users and their applications
- Overall throughput performance required for the application environment and estimates for the number of concurrent sequential I/O streams
- Growth factors for future data storage and performance

Workflow Requirements

There are many application types, use cases, and resulting I/O patterns in NetApp HPC Solution for Lustre environments. For sizing purposes, some simplifying assumptions are made about the characteristics of parallel I/O that are relevant for most HPC environments. In general, the highest throughput performance is required when clients of a cluster are executing large parallel sequential file read and write requests from the Lustre file system. These sequential file reads and writes are referred to as streams.

To the storage elements in the NetApp HPC Solution for Lustre, this application I/O typically appears as many evenly distributed, concurrent, sequential read and write streams spread across the storage systems. The NetApp HPC Solution for Lustre has been measured under similar test conditions to characterize the throughput performance across a wide range of concurrent, sequential I/O streams for reads and writes of various I/O sizes.

Capacity Requirements

For either existing applications usage or new application deployment, it is necessary to determine the amount of storage capacity required across the file system. The storage capacity is usually specified in terabytes. Also consider the number of users and the total number of files in use.

Performance Requirements

Performance is usually specified as the total storage throughput required at peak periods. This is measured in both new and existing application environments as the intended gigabytes per second in total for data access across all applications.

Also important for performance is the number of concurrent streams that are executed to achieve the overall throughput. It is necessary to use both total throughput and the number of streams in calculations for sizing. Also of note are the percentages relative to read and write for the total throughput requirements. Currently, however, sizing is calculated only for the throughput required for the aggregate of read and write streams.

Metadata performance has also been characterized for operations such as:

- Number of file creates and deletes per second
- Directory creates and deletes per second
- Number of stats per second (operations such as the `ls` command in Linux)

The metadata performance is not constrained by the performance of the E-Series storage array and is mostly dictated by the choice of the MDS and by the Lustre architecture.

Calculation of Storage Sizing Requirements for Lustre

Two sizing calculations should be performed for Lustre file systems:

- Storage for use as OSTs
- Storage for use as MDTs

The OST storage capacity requirements are specified by customers for their particular environments. The MDT storage capacity requirements are determined by the total number of files that are required by the customer across the Lustre file system.

Note: The calculations for Lustre storage requirements for this document are based on a single Lustre file system.

Lustre Metadata

For E-Series storage for the Lustre MDT, NetApp recommends the E2624, which is configured to provide maximum metadata performance and satisfy capacity requirements. A single E2624 shelf with up to 24 600GB hard-disk drives (HDDs) provides enough capacity to support petabyte-sized file systems (depending on average file size) and meets the storage performance requirements such that metadata performance is limited by the MDS hardware performance, not by the E-Series storage performance.

Calculating Metadata Storage Requirements

In calculating the MDT size, the important factor to consider is the number of files to be stored in the file system. This number determines the number of inodes needed, which drives the MDT sizing. Attached storage required for Lustre metadata is typically 1% to 2% of the file system capacity, depending upon file size.

Assuming the default Lustre value of 4KB per inode, the MDT storage capacity requirement is calculated as follows:

- Required capacity = total number of files in the file system x 4KB per inode

If the total number of files is not known, that number can be estimated by dividing the total file system capacity by the average file size. For example, if the average file size is 5MB and there are 500TB of usable OST space, then the estimated maximum number of files (and therefore the minimum number of inodes) can be calculated as follows:

- (500TB x 1024GB/TB x 1024MB/GB) / 5MB per inode = 104.9 million inodes

NetApp recommends using at least twice the minimum number of inodes to allow for future expansion and for an average file size that is smaller than expected. Thus, the required space is:

- 4KB per inode x 209.8 million inodes = 839.2GB

If the average file size is small (for example, 4KB), Lustre is not very efficient because the MDT uses as much space as the OSTs; however, this is not a common configuration for Lustre.

Also, if the MDT is too small, this can result in unusable or inaccessible space on the OSTs. Before formatting the file system, be sure to determine the appropriate size of the MDT that is needed to support the file system. It is difficult to increase the number of inodes after the file system is formatted.

The E2624 with 24 600GB HDDs is the default metadata storage supported for the NetApp HPC Solution for Lustre.

Best Practice
For RAID configuration of this array for metadata, use RAID 10 (11+11) drives with 128KB segment size and two hot spare drives.

Given this configuration, Table 3 specifies the maximum capacity of the E2624.

Table 3) Drive size and capacity.

E-Series	Controller Shelf	Drives	Drive Sizes	Raw Capacity	Formatted Capacity (RAID 10)
E2600	E2624	24	600GB	14.4TB	6.0TB

Note: The recommended configuration has two hot spare drives in the Lustre E2624 MDT. This configuration limits the usable drives to 22, out of which 50% is usable capacity because of RAID 10 overhead. All 24 drives can be used for 6.5TB of usable space, but without the added data protection.

The Lustre file system can support a maximum of 4 billion files. This is equivalent to 4 billion inodes. As a result, for very large Lustre file system deployments, additional DE5600 disk expansion shelves may be required to provide the number of required inodes.

Metadata Server Recommendation

NetApp recommends using the fastest CPUs for the MDS. For performance reasons, faster CPUs minimize the impact of large numbers of client locks. It is also advantageous to use fewer faster cores rather than many slower cores.

NetApp also recommends using dedicated system disks (RAID 1 or 10) for the operating system (OS), separate from the Lustre file system (MDT), which is placed on the E-Series storage.

Calculating Metadata Server Memory Requirements

MDS memory requirements might depend on a number of factors, such as number of clients, size of the directories, metadata load placed on the server, and so on.

The amount of memory used by the MDS depends on how many clients are supported by the system and how many files they use in their working sets. Available memory is consumed by storing the file metadata and the required locks for a given client. The number of locks held by clients varies by load and memory availability on the server. A client can hold in excess of 10,000 locks at times. MDS memory usage is roughly 2KB per file, including the Lustre distributed lock manager (DLM) lock and kernel data structures for the files currently in use. Caching file data can improve metadata performance by a factor of 10 or more compared to reading it from disk. By default, 400MB are used for the file system journal. Additional RAM is used for caching file data for the larger working set, which is not actively in use by clients but which should be kept “hot” for improved access times.

In short, MDS memory requirements include these considerations:

- File system metadata requires a reasonable amount of RAM for file system metadata workflows. Although no hard limit can be placed on the amount of file system metadata, if more RAM is available, then disk I/O is needed less often to retrieve the metadata.
- Network transport workflows also use host memory associated with TCP or other network protocols that use send/receive buffers. This memory requirement must also be taken into consideration.
- Journal size by default is 400MB for each Lustre file (`ldiskfs`) system and can use an equal amount of RAM on the MDS node for each file system.
- In a failover configuration, the secondary MDS must be equipped with at least the same amount of memory as the primary MDS.
- Memory usage includes the memory needed for the Lustre file system, but it also includes the requirements from the OS and from the resident applications sitting on top of the OS. This can require additional gigabytes of RAM.

Having additional memory available might significantly improve performance. For directories containing one million or more files, more memory might provide a significant benefit. For example, in an environment in which clients randomly access one of 10 million files, having extra memory for the cache

significantly improves performance. For the MDS, NetApp recommends a minimum of 24GB of memory, whereas for improved performance, configuring the MDS with 32GB to 48GB of memory would be beneficial.

Estimating Metadata Performance

A single-shelf E2624 with 24 HDDs is used for the MDT storage in the NetApp HPC Solution for Lustre. With this configuration and proper settings, the metadata performance is limited by the choice of server for the MGS, not by storage performance of the E2624. For sizing purposes then, the only required variable is storage capacity because performance is not a factor.

Metadata performance was tested with a dual-socket Westmere-class server with 48GB of memory as the MDS, and an E2624 with 24 2.5" SAS 600GB 10K RPM HDDs in the recommended configuration for the MDT performance of RAID 10 (11+11) with 128KB segment size. Table 4 lists the representative ranges of performance for metadata operations that can be expected with this configuration.

Table 4) Metadata performance.

Operation	Performance (Operations per Second)
Directory creation	6,000 to 23,000
Directory stat	10,000 to 89,000
Directory deletes	3,700 to 12,000
File creation	3,900 to 20,000
File stat	3,800 to 81,000
File deletes	5,600 to 21,000

Metadata performance numbers vary, depending on the number of threads and the structure of the underlying jobs being run on Lustre, which produce varying file sizes and numbers of files per directory.

Lustre Object Storage

Sizing E-Series storage for the Lustre OST requirements begins with the dual requirements of capacity and performance. For OST purposes, there are two platform choices:

- E5460 with 3.5" 7.2K RPM HDDs
- E5424 with 2.5" 10K RPM HDDs

Each platform provides specific performance, capacity, and scale-out capabilities.

This guidance assumes that customers have already selected the basic drive technology that they want to use, either 3.5" SAS 7.2K RPM or 2.5" SAS 10K RPM HDDs. The tradeoffs between these drive types include form factor, capacity per disk, performance, and other features. The selection guidance for drive technology is beyond the scope of this document, but sizing rules are provided for both.

Sizing for performance is done first, and it determines the number of controller shelves needed to meet the requirement. Capacity is then calculated for the resulting controller shelves required for performance. If this capacity does not meet the overall requirement, extra expansion shelves are added to scale out to the required capacity.

This guidance also assumes linear performance scaling as OSSs and OSTs are added to the file system. Although this is a reasonable assumption for Lustre in general, many issues might prevent a cluster of client servers from achieving 100% linear performance scaling with additional OSSs and OSTs. These issues include the client cluster network design and performance, the distribution of files and workloads across those OSSs and OSTs, and varying approaches to determining how applications perform I/O and

interact with the parallel file systems. These issues are beyond the scope of this document but are the subject of future work planned for this solution.

Object Storage Server Recommendation

The OSS can use any modern X86_64 CPUs. For maximum performance, use as much RAM as possible. Consider that there might be multiple OSTs mapped to each OSS, each of which requires file system journal space, memory for each I/O thread, and OST cache for each connected OST. Spare memory is also used for the OSS read cache. In general, OSS RAM requirements are higher than MDS requirements, although the CPUs used might not be as important. The system OS should also reside on a physically separate partition from the E-Series storage used for OSTs.

Best Practice
NetApp recommends using dedicated system disks (RAID 1 or 10) for the OS, separate from the Lustre file system OST, which is placed on the E-Series storage.

Sizing for Object Storage Performance

The storage performance sizing requirement has two components:

- Total storage I/O bandwidth, or throughput, required across the entire Lustre file system (usually stated in gigabytes per second)
- Maximum number of concurrent I/O streams from all clients across the entire file system to achieve the required total aggregate throughput

Testing was developed to characterize the NetApp HPC Solution for Lustre storage performance and to create metrics for sizing calculations that are representative of customer use cases and environments. Most customer environments for the NetApp HPC Solution for Lustre present storage workloads including concurrent file read and write streams that, in aggregate, define the required system throughput. Testing produced a synthetic workload with ranges of concurrent I/O streams and I/O sizes to simulate the HPC environment.

From this testing, two operating regions were identified and metrics developed for sizing. In the first region (region 1), testing discovered that maximum saturated performance for the E-Series controllers was achieved for small to moderate numbers of concurrent streams. This is the preferred operating region for environments that require the highest performance. Application workloads are constrained to operate below the maximum number of supported concurrent streams for this region to achieve this performance metric.

As the number of concurrent streams is increased, the streams intermix and present a more random than sequential I/O workload to the storage. In region 1, the Lustre OSS and E-Series controllers are able to process the intermixed concurrent I/O streams and still achieve the saturated throughput of the controllers. Above the maximum stream count for region 1, the I/O randomness results in drive-limited performance. This is the region 2 performance metric for workloads that required large numbers of concurrent I/O streams.

Sizing for performance is then determined by the total aggregate throughput required and the total number of concurrent I/O streams at that throughput. These two values produce work characteristics that are better met in either region 1 or region 2. Depending on the total number of concurrent streams required, either region 1 (moderate stream counts) or region 2 (large stream counts) produces the optimal number of controller shelves.

Sizing for these two regions is calculated in the following way:

- Inputs:
  - Controller shelf type is either E5460 or E5424.

  - MaxThroughput is defined as aggregate throughput required in gigabytes per second.
  - MaxStreams is defined as total concurrent I/O streams at aggregate throughput.
- Outputs:
  - Required controller shelves of either type E5460 or E5424
  - Minimum drive and shelf requirements to achieve the required performance
- Calculations:
  - Total controller shelves required = either [region 1 controller shelves or region 2 controller shelves]
  - Region 1 controller shelves = greater of [MaxThroughput / (region 1 controller shelf performance for the array type) or MaxStreams / (region 1 max streams)]
  - Region 2 controller shelves = greater of [MaxThroughput / (region 2 controller shelf performance for the array type) or MaxStreams / (region 2 max streams)]

Table 5 provides performance sizing metrics for region 1 (a moderate number of streams) and region 2 (a high number of streams).

Table 5) Performance metrics for region 1 and region 2.

Array Model	Number of Drives	RAID Format	Region 1: Maximum Streams	Region 1: Controller Shelf Performance (GB/Sec)	Region 2: Maximum Streams	Region 2: Controller Shelf Performance (GB/Sec)
E5460	30	RAID 6 (8+2)	50	1.25	1,500	0.6
E5460	60 to 360	RAID 6 (8+2)	100	2.5	3,000	1.2
E5424	24	RAID 6 (8+2)	50	1.3	1,000	0.6
E5424	48 to 192	RAID 6 (8+2)	100	2.6	2,000	1.2

Performance is sensitive to the record size used for the reads and writes. Although 1MB is the recommended record size for the Lustre file system, record sizes between 512KB and 4MB might still meet or exceed the performance results given in the sizing metrics, depending on how the clients aggregate the data communication. This behavior was observed during performance testing.

Sizing for Object Storage Capacity

When sizing for performance is complete, the resulting capacity can be calculated and compared against the overall capacity requirement. If the resulting capacity is insufficient, additional expansion shelves with drives can be added based on the extra capacity necessary to meet the requirement.

Best Practice
NetApp recommends using the configuration described in this document: RAID 6 (8+2) with 128KB segment size for a total RAID stripe size of 1MB. This is the best fit for Lustre file systems, which are optimized around a 1MB SCSI block level I/O size. NetApp recommends this configuration for both of the drive technologies and for both the E5460 and the E5424 storage arrays.

The E5460 supports either three or six RAID 6 (8+2) volume groups, using either a half-populated (30-drive) shelf or a fully populated (60-drive) shelf. This is the same configuration used for expansion with the DE6600 expansion shelf.

The E5424 supports two RAID 6 (8+2) volume groups (for a total of 20 drives) with up to four hot spare drives per shelf. This is the same configuration used for expansion with the DE5600 expansion shelf.

Table 6 lists capacity sizes for various HDD and drive shelf options.

Table 6) Drive and shelf capacities.

E-Series Controller	Controller Shelf	Expansion Shelf	Drives	Drive Sizes RAID	Raw (TB)	Formatted (TB) RAID 6 (8+2)
E5400	E5460 (half populated)		30	2/3TB	60/90	43.7/65.5
ESM		DE6600 (half populated)	30	2/3TB	60/90	43.7/65.5
E5400	E5460		60	2/3TB	120/180	87.3/130.9
ESM		DE6600	60	2/3TB	120/180	87.3/130.9
E5400	E5424		24	600GB/900GB	14.4/21.5	8.7/13.1
ESM		DE5600	24	600GB/900GB	14.4/21.5	8.7/13.1

Note: Only two RAID 6 (8+2) volume groups are specified for the E5424 with the Lustre file system. This limits the usable drives to 20, in which 80% is usable capacity because of RAID overhead. The remaining four drives are available for hot spare drives.

When configuring expansion shelves (DE6600, DE5600), it is important to observe the maximum drive counts per array in order to not exceed allowable limits for the respective controllers. Table 7 provides the maximum drive capacity per shelf.

Table 7) Maximum drive capacity per shelf.

E-Series Controller	Controller Shelf	Expansion Shelf	Maximum Number of Drives	Maximum Number of Expansion Shelves Supported
E5400	E5460 with CE6600 shelf		60 (360 total per array max.)	5 (6 total shelves per array max.)
ESM		DE6600 shelf	300	5
E5400	E5424 with CE5600 shelf		24 (192 total per array max.)	7 (8 total shelves per array max.)
ESM		DE5600 shelf	168	7
E2600	E2624 with CE5600 shelf		24 (192 total per array max.)	7 (8 total shelves per array max.)
ESM		DE5600 shelf	168	7

It is possible to mix expansion shelves (DE6600, DE5600), with the rule that the maximum number of drives cannot exceed the limit for each controller type. Typically, however, NetApp does not recommend mixing shelf and drive types with the NetApp HPC Solution for Lustre because this solution favors symmetrical designs. Any use of a mixed shelf configuration requires validation testing to verify that sizing and performance criteria are achievable.

When expansion shelves or drive count requirement maximums exceed the limit for a controller shelf, it is necessary to configure additional E-Series arrays to meet the desired capacity requirements.

Once the sizing for performance is completed and the number of controller shelves with drives is calculated, the capacity for this performance storage is calculated according to Table 7. If the overall capacity required exceeds the capacity achieved through performance sizing, additional drive shelves

and drives can be added to achieve the overall capacity. The total number of expansion shelves is limited as shown in Table 7.

Each time a RAID 6 (8+2) volume group is added, it is mounted as an OST by Lustre. Additional OSTs added through expansion drive shelves should be allocated evenly across each array to maintain symmetrical OSS performance and to achieve symmetrical file system capacity across the OSSs and the entire file system. In the end, each array should consist of an identical number, type, and capacity of OSTs.

If the overall capacity cannot be achieved after reaching the maximum number of expansion shelves behind each controller shelf, additional controller shelves with symmetrical expansion shelves must be added to reach the overall required capacity.

Sizing Example

In this example, a customer needs a Lustre file system configuration that supports 8.4GB/sec write I/O throughput for 300 concurrent I/O streams, with a total user storage capacity of 1PB.

Using the E5460, the total number of shelves is:

- 1024TB/PB x 1PB / 130.9TB per shelf = 7.82 total shelves.

For 8.4GB/sec with 300 concurrent streams, four controller shelves are required:

- (8.4GB/sec / 2.5GB/sec per controller for region 1) = 4 (rounded up)
- (300 streams / 100 streams per controller for region 1) = 3

Required controller shelves for performance = greater of throughput or streams = four controller shelves.

Using the table for formatted capacity with 3TB drives in the E5460, the capacity for the four controller shelves is:

- Capacity for controller shelves = 130.9TB x 4 = 523.6TB
- Expansion capacity required = 1000TB – 523.6TB = 476.4TB
- Expansion capacity recommendation = (476.4TB / 139.9TB per DE6600) = 4 DE6600 expansion shelves (rounded up)

Given the requirement to keep the design symmetrical, the expansion capacity needed is one DE6600 expansion shelf with 60 3TB drives behind each controller shelf (eight shelves total).

The configuration summary is four E5460 systems with four additional DE6600 expansion shelves for a total of 480 3TB drives, which provides an overall performance of 10GB/sec and 1047.2TB of capacity.

Scale-Out Capacity and Performance Sizing

Sizing Lustre storage capacity and performance involves calculating the necessary number of controller shelves and expansion shelves for the OSTs. This guidance assumes linear performance scaling as OSTs and OSSs are added. Likewise, linear capacity scaling is achieved as controller shelves and expansion shelves are added. This straightforward scale-out methodology is a key attribute of the NetApp HPC Solution for Lustre.

Summary

The E-Series storage systems platforms provide the storage capacity and performance needed to meet Lustre application requirements.

2.2 NetApp HPC Solution for Lustre Performance Considerations

Overview

Performance is an essential component of the NetApp HPC Solution for Lustre. Understanding the typical workloads for NetApp HPC Lustre environments and how to maximize the NetApp HPC Solution for Lustre for optimal performance for those workloads is critical.

Key to characterizing performance and enabling optimization is building a test methodology that is representative of real-world workloads. Although there are many different and varied HPC applications, these applications have some common requirements from the parallel storage environment. Clustered supercomputers run application codes that generally must perform parallel reads and writes from the file system into and out of the memories of the client nodes.

Although there are different strategies for how these applications store information on the file system, a general approach for achieving optimal parallel I/O performance is for each process on the supercomputer to write or read its information using its own dedicated single file. This minimizes contention between nodes and processes for reading and writing files in parallel and allows for symmetrically distributing I/O streams across the entire storage system for maximum throughput.

Because each of these processes typically opens a file and performs a complete read or write, the resulting I/O stream is 100% sequential read or write, and the Lustre file system client aggregates requests and tries to produce aligned, large, well-formed I/O for best throughput.

Another strategy is for all processes to read and write from one large single file. This approach is usually less efficient because of file locking and the complexities of striping read and write I/O streams across every parallel file server. However, applications that take this approach mitigate the file-sharing and locking problem by allocating portions of the file space to specific processes, and Lustre provides locking granularity that allows those processes to read or write their portions exclusively. Therefore, the test methodology used for this characterization is still applicable.

Performance Characterization for Lustre

Two aspects of E-Series storage performance must be measured and characterized for Lustre:

- Performance for file data path I/O to OSTs and metadata operations
- Performance using MDTs

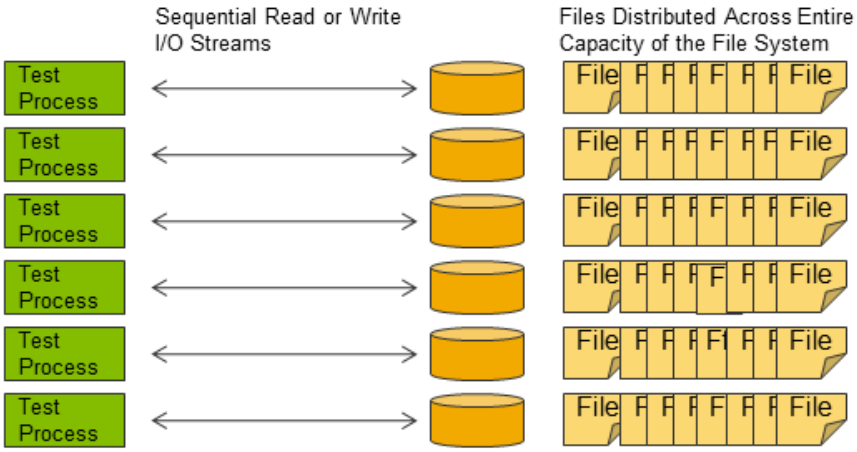
The methodology for testing OST performance characterizes parallel file system throughput using volumes on E5460 or E5424 OST storage. The requirement for OST performance is for optimal large I/O sequential throughput from multiple concurrent streams.

The methodology for testing metadata performance characterizes the input/output operations per second (IOPS) behavior of the MDS, using volumes on the E2624 as MDT storage. Metadata storage operations are typically random small block I/O, and they require optimal IOPS performance from the MDT storage configuration.

Lustre Object Storage Target Storage Performance

The test methodology that was chosen for characterizing and measuring solution performance emulates the process-per-file strategy. A multiserver, file-based I/O benchmarking tool was used to generate workflow models ranging from file read and write streams to files configured across the entire capacity of the file system. Test tools such as vdbench and iozone were used to provide the ability to format files on the file system and distribute file read and write processes across a number of test client nodes, synchronize the I/O activities across these nodes, and coalesce results into summary data files. Figure 4 illustrates this methodology.

Figure 4) Test process I/O streaming to the file system.



The methodology for testing file read and write I/O performance across the file system includes the following tasks:

- Format thousands of files of sufficient size to use the entire file system capacity and therefore to span all sectors of all drives in the test file system.
- Run a series of automated test scripts on a number of client nodes that measure throughput with a number of processes, each reading or writing sequential streams to individual files across the file system.
- Run these tests for ranges from 6 to 1,000s of total streams per E-Series storage array.
- At each test case for a number of streams, run various I/O sizes and run from 100% reads to 100% writes, as well as combinations in between.

During testing, these automated tests were run across a range of storage configurations and across a range of numbers of OSTs for a single E5460 or E5424 controller shelf with a range of drive counts and expansion shelves. The aggregate throughput results for both reads and writes were plotted against the number of streams to determine the performance behavior for a range of file-per-process I/O streams that HPC applications might provide.

Figure 5 shows a qualitative view of this behavior for 100% read streams, 100% write streams, mixtures of read and write streams, and various I/O sizes.

Figure 5) Performance regions by stream counts.

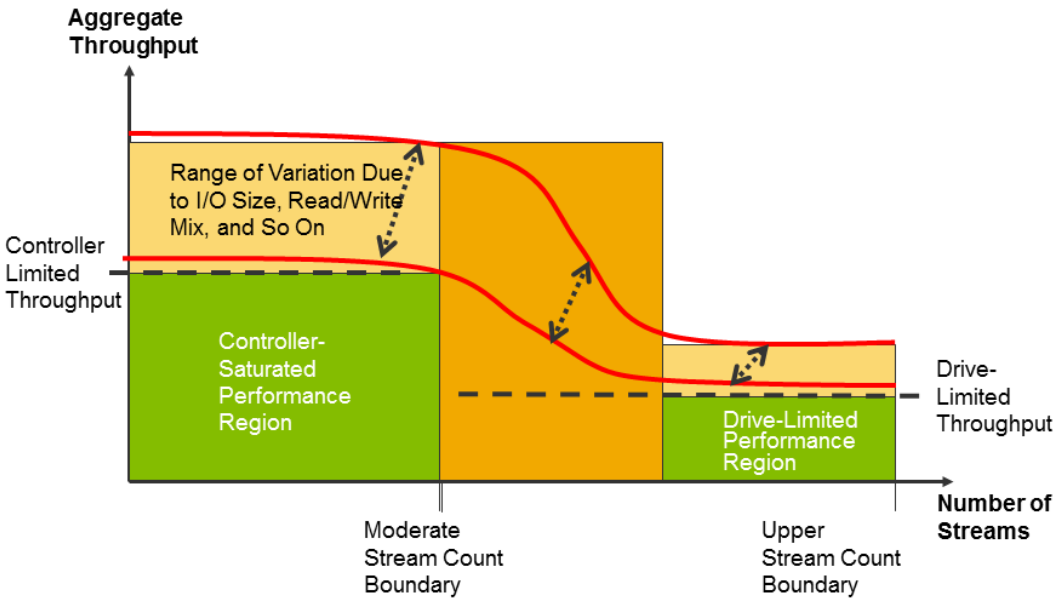


Figure 5 shows aggregate throughput versus stream count for a single E-Series controller shelf, with bands of values for various I/O sizes and combinations of read and write streams. This is the view of performance from the perspective of applications on the clients making file read and write system calls on the client server nodes. The regions indicate the overall response of the Lustre file system client, the Lustre network, the OSSs, and the OST storage mounted by those OSSs. This provides an integrated view of system throughput, which is similar to the throughput of an actual application performing parallel sequential file reads and writes.

The first green region demonstrates the value and efficiency of the NetApp HPC Solution for Lustre by showing that HPC applications can yield controller-saturated performance for low to moderate stream counts for a variety of I/O sizes and mixtures of read and write streams. This is achieved because buffers in the Lustre client (OSS) and caches in the E-Series RAID controllers are able to aggregate and reorder I/O sufficiently to provide optimal, well-formed I/O requests to the disk hard drives in the E-Series systems. This well-formed I/O to the disks means that the HDDs are used in streaming mode and can keep pace with the throughput of the controllers. This is important in sizing because it simplifies potentially complex customer architectures by providing a flat performance level that can be achieved for each controller shelf over a range of conditions. Total aggregate performance can then be achieved by adding controller shelves through linear throughput scale-out, using the Lustre file system.

The green controller-limited throughput region is determined by selecting the test result for the condition that creates the minimum saturated controller-limited throughput, adjusted to account for test and manufacturing variations. This creates a high-confidence sizing number that works for a wide range of conditions and provides the value used in the sizing algorithm.

In Figure 5, the first yellow region above the controller-saturated green region shows performance results that can be achieved through optimizations such as selecting specific read/write mix percentages or specific I/O sizes. Customer designs may plan for sizing in this region if specific conditions are met, but there is increased risk that performance could vary and might not meet requirements. NetApp recommends doing proof-of-concept testing of actual environments if sizing in the yellow regions is required.

As stream counts increase past the moderate stream count boundary, the mixture of larger numbers of concurrent I/O streams spreads widely across the address space, making it increasingly difficult to keep the HDDs in fully streaming mode. As stream counts continue to increase to the upper stream count

boundary, the aggregate throughput becomes drive limited and represents the large I/O random performance of the HDDs in the E-Series system. For sizing purposes, it is advisable to use large-stream-count sizing rules in the darker yellow region, where throughput is transitioning from controller-saturated to drive-limited performance.

In the second green region in Figure 5, the aggregate throughput is determined by selecting the test result for the condition that produces the minimum drive-limited throughput, adjusted to account for test and manufacturing variations. Sizing rules in this region set the aggregate throughput to the drive-limited value for stream counts from the moderate stream count limit to the upper stream count limit.

The second yellow region corresponds to variations above the drive-limited throughput caused by mixes of I/O sizes and reads versus writes. Again, if sizing optimization requires operating in this region, a proof-of-concept test is required in order to have confidence in the expected performance.

E5460 Performance Test

The E5460 storage system was used for Lustre OSTs with a full complement of drives. Two OSSs were connected to the E5460 through two quad data rate (QDR) (40Gb) IB ports per OSS and were cross connected to the two controllers in the E5460. These OSS servers were connected to a Lustre client network using QDR IB.

Test Environment

For the Lustre OSS/OST performance metrics, the E5460 was configured with 60 3.5” 3TB SAS drives in 10-drive RAID 6 (8+2) volume groups with 128KB segment sizes. There was one volume per volume group for maximum performance, thus six RAID 6 (8+2) volumes per E5460.

Results and Analysis

The minimum aggregate read and write throughput for the E5460 with six RAID 6 (8+2) LUNs used as OSTs was 2.5GB/sec for 100 streams or less. Based on test results, 100% reads are specified at 3.4GB/sec, and 100% writes are specified at 2.5GB/sec.

For stream counts from 100 to 3,000, the minimum aggregate bandwidth was measured to be no less than 1.2GB/sec.

A half-shelf configuration (30 3.5" 3TB drives) was also tested. For 50 streams or less, the aggregate bandwidth was measured at no less than 1.25GB/sec, and for 50 to 1,500 streams, this value was no less than 0.6GB/sec.

Adding expansion shelves and more than 60 drives did not increase the controller-saturated performance results significantly. It is more effective to achieve higher performance by adding controller shelves. As a result, NetApp does not recommend adding expansion shelves to achieve performance gains when throughput performance is the main objective. Although NetApp recommends expansion shelves for adding capacity, little additional performance gain was observed during testing.

Conclusion

Each E5460 controller shelf (60 3TB drives) provides a minimum of 2.5GB/sec of aggregate read and write bandwidth for all combinations of read and write percentages of sequential streams from 6 to 100 streams per stripe group. In order for these results to hold valid, the workload stream must be capable of driving the data.

Each E5460 controller shelf (60 3TB drives) provides a minimum of 1.2GB/sec of aggregate read and write bandwidth for all combinations of read and write percentages of sequential streams for anywhere from 100 to 3,000 streams per stripe group.

E5424 Performance Test

The E5424 storage system may also be used for Lustre OST for use with 2.5" SAS drive technology. Optionally, customers can use this technology if they do not require the additional capacity provided by 3.5" drives and if they want the higher drive performance, smaller footprint, and lower power consumption associated with the 2.5" drives.

Test Environment

For the Lustre OSS/OST performance metrics, the E5424 was configured with 24 2.5" 600GB SAS drives running at 10K RPM. These drives were configured in two sets of 10-drive RAID 6 (8+2) volumes with four hot spares. There was one volume per volume group for maximum performance.

Results and Analysis

The minimum aggregate read and write throughput for the E5424 (2 RAID 6 [8+2] LUNs) was measured at 1.3GB/sec for 50 streams or less. Based on test results, 100% reads are specified at 1.3GB/sec, and 100% writes are specified at 1.6GB/sec.

For stream counts from 50 to 1,500, with two LUNs in the E5424 controller shelf, the aggregate read and write throughput was no less than 0.6GB/sec.

Doubling the number of drives with the use of an E5424+DE5600 (48 drives total) configured with four RAID 6 (8+2) LUNs essentially doubled the performance metrics. The minimum aggregate read and write throughput for the E5424+DE5600 (4 RAID 6 [8+2] LUNs) was 2.6GB/sec for 100 streams or less. Based on test results, 100% reads are specified at 2.6GB/sec, and 100% writes are specified at 2.6GB/sec.

For stream counts from 100 to 2,000 with four LUNs, the aggregate read and write throughput was no less than 1.2GB/sec.

Adding expansion shelves beyond the controller shelf (48 drives total) did not increase the controller-saturated performance results significantly. It is more effective to achieve higher performance by adding controller shelves. As a result, NetApp does not recommend adding expansion shelves (beyond the first 24-disk expansion shelf) to achieve performance gains when throughput performance is the main objective. Although NetApp recommends expansion shelves and additional drives for adding capacity, in the testing for sizing purposes, little additional performance gain was observed when adding more than one expansion shelf.

Conclusion

Each 24-drive E5424 controller shelf provides 1.3GB/sec of aggregate read and write bandwidth for all combinations of read and write streams from 1 to 50 per stripe group. Doubling the drive count (that is, providing a second drive shelf) to use 48 drives provides a minimum aggregate read and write bandwidth of 2.6GB/sec.

Summary of Performance and Sizing Guidelines

Table 8 summarizes the performance and sizing guidelines for the E5460 and E5424 based on the number of drives, the number of concurrent streams, and the resulting performance. It includes numbers for region 1 (a moderate number of streams) and region 2 (a high number of streams).

Table 8) Performance summary by array.

Array Model	Number of Drives	RAID Format	Region 1: Maximum Streams	Region 1: Controller Shelf Performance (GB/Sec)	Region 2: Maximum Streams	Region 2: Controller Shelf Performance (GB/Sec)
E5460	30	RAID 6 (8+2)	50	1.25	1,500	0.6
E5460	60 to 360	RAID 6 (8+2)	100	2.5	3,000	1.2
E5424	24	RAID 6 (8+2)	50	1.3	1,000	0.6
E5424	48 to 192	RAID 6 (8+2)	100	2.6	2,000	1.2

Lustre Metadata Performance

Metadata storage is provided by the E2624 with 600GB or 900GB SAS drives that run at 10K RPM. These are configured into RAID 10 LUNs, which are then presented to Lustre MDSs for use as Lustre MDTs.

Metadata performance is primarily a function of the performance of the MDS. When configured correctly, it is not limited by the E2624 MDT storage. The performance of metadata operations is characterized here to confirm that the storage is not the limiting factor and to provide some guidance for performance expectations using a typical MDS server.

Test Environment

For the Lustre MDS/MDT performance metrics, the E2624 was configured with 24 600GB drives running at 10K RPM. The RAID configuration was RAID 10, and all 24 drives were used in a single volume group with a single volume used as the Lustre MDT.

Four dual-socket Westmere (E5620) servers were used as the client nodes, two dual-socket Westmere (E5620) servers as the OSS nodes, and one dual-socket Westmere (E5620) server as the MDS. All servers had 48GB of RAM. The metadata performance benchmark tool used was `mdtest`. It was compiled with `mvapich` 1.2.0. `mvapich` 2 1.4 was also tested, but the benchmark data did not warrant its widespread use for these tests.

Performance Test Results

Figure 6 illustrates `mdtest` performance test results for directory operations using 1,000 files per directory.

Figure 6) Directory operations performance.

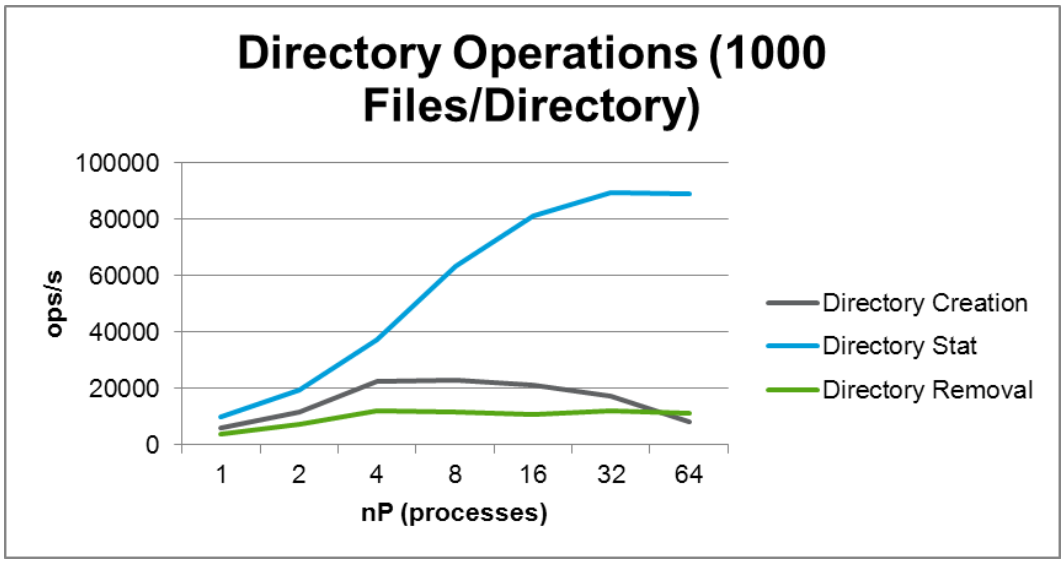
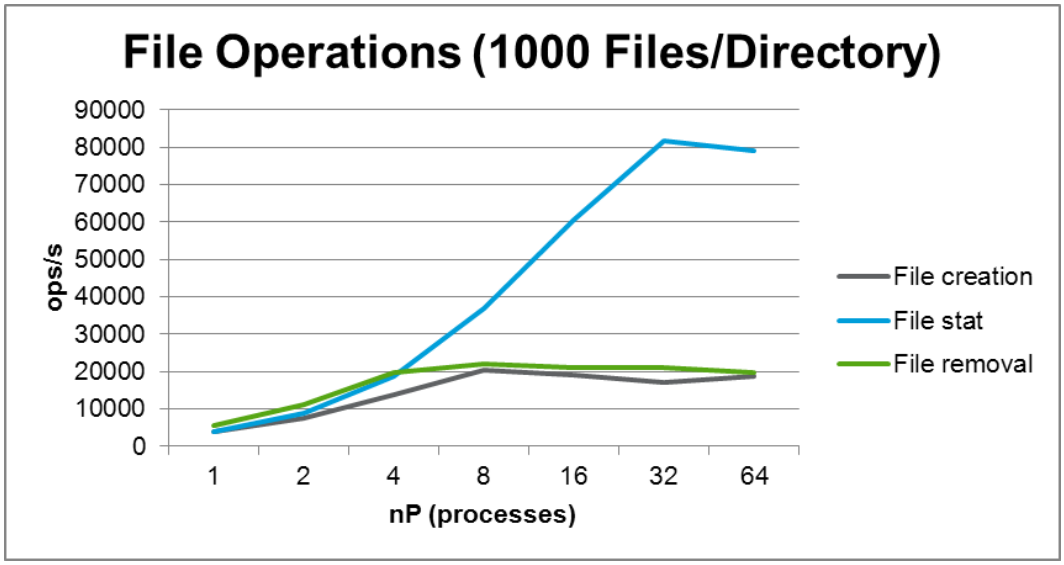


Figure 7 illustrates `mdtest` performance test results for file operations using 1,000 files per directory.

Figure 7) File operations performance.



The test data shows that directory creations peaked with at least 23,000 operations/sec, directory statistics at 89,000 operations/sec, and directory removals at 12,000 operations/sec. There was a peak of at least 20,000 file creations/sec, 81,000 file statistics/sec, and 21,000 file deletions/sec.

Different hardware with different message passing interface (MPI) versions and server/client settings will produce different test results. Various test parameters, such as different numbers of clients initiating metadata operations and creating or deleting different numbers of files per directory, will produce higher or lower performance numbers. Therefore, these performance numbers should not be used as absolute metadata MDS performance numbers but rather as a guideline for the metadata performance that is achievable using the given hardware.

Analysis

Testing was done with `mdtest`, using the recommended segment size of 128KB and all 24 drives in an E2624. The 24 (600GB) drives were placed into a single RAID 10 group (12+12). Essentially, `mdtest` was run in a loop using four clients, two OSSs, six OSTs, and a single MDS/E2624 MDT. The following run script was used:

```
mpirun -np $NP -hostfile hostfile.16 ./mdtest -n 1000 -i 1 -v -u -d /mnt/lustre/
```

Where:

- `NP` is the number of MPI processes distributed evenly among the clients, where possible
- `hostfile` is the hostfile for all of the machines with as many lines as tests to run
- `n` is the number of files per directory
- `i` is the iteration number
- `v` is for verbose output
- `u` creates a unique working directory for each task
- `d` is the working directory in which to run the tests

Conclusion

The test data indicates that the E2624 supports directory creations of at least 23,000 operations/sec, directory statistics of at least 89,000 operations/sec, and directory removals of at least 12,000 operations/sec. There is a demonstrated minimum of 20,000 file creations/sec, 81,000 file statistics/sec, and 21,000 file deletions/sec. Higher results are expected as the number of clients and OSTs increases and the testing process is broadened.

Scale-Out Performance Considerations

For Lustre metadata, scale-out includes the expansion of the MDT, using additional RAID 10 drive pairs in expansion shelves up to the 192 drive-count limit of the E2624. If it becomes necessary to scale beyond the IOPS capability of hard disks, IOPS availability for metadata use can be increased through the use of solid-state drives.

For Lustre OST storage, the E5460 and the E5424 storage systems may be expanded using expansion shelves to meet capacity and throughput requirements. Capacity requirements are met through the use of multiple shelves with the appropriate drives for the E5460 or the E5424. The number of controller shelves needed depends on the throughput and stream count requirements.

Summary

The E-Series E2624, E5424, and E5460 storage systems provide the necessary capacity, throughput, IOPS, and response times to meet performance requirements for demanding Lustre application environments.

2.3 E-Series Solutions Hardware Packaging

Table 9 lists the part numbers associated with all E-Series solutions. These are the part numbers that are included in the EzChoice Quote Tool.

Table 9) E-Series part numbers.

Category	Part Number	Product Description
System enclosures	DE6600-SYS-ENCL-R6	Enclosure, 4U-60, DE6600, empty, 2PS
	DE5600-SYS-ENCL-R6	Enclosure, 2U-24, DE5600, empty, 2PS
	DE1600-SYS-ENCL-R6	Enclosure, 2U-12, DE1600, empty, 2PS
Expansion enclosures	E-X5680A-QS-R6	Enclosure, 4U-60, DE6600, empty, 2PS, QS
	E-X5681A-QS-R6	Enclosure, 2U-24, DE5600, empty, 2PS, QS
	E-X5682A-QS-R6	Enclosure, 2U-12, DE1600, empty, 2PS, QS
ESM controller	E-X30030A-R6	ESM controller, SBB-2
5400 controllers	E5400A-12GB-R6	E5400A, 12GB controller
	E5400A-6GB-R6	E5400A, 6GB controller
2600 controller	E2600A-2GB-R6	E2600A, 2GB controller
Host interface cards (HICs)	X-52708-00-R6	HIC, E5400, 40GB, IB, 2-port
	X-48855-00-R6	HIC, FC, 4-port, 8Gb, E5400
	X-52709-00-R6	HIC, E2600, 1GB iSCSI, 4-port
	X-52710-00-R6	HIC, E2600, 10GB iSCSI, 2-port
	X-52194-00-R6	HIC, E2600, FC, 4-port, 8Gb
	X-52195-00-R6	HIC, E2600, SAS, 2-port, 6Gb
Racks	X-M102061-R6	40U rack, empty, L6-30, domestic
	X-M102062-R6	40U rack, empty, IEC309, international
Software (point of sale)	SW-5400-FDE-SKM-P	SW, full-disk encryption (FDE) security key management, 5400, -P
	SW-5400-SNAPSHOT-P	SW, Snapshot™, 5400, -P
	SW-5400-VOL-COPY-P	SW, volume copy, 5400, -P
	SW-5400-REM-MIRR-P	SW, remote mirroring, 5400, -P

Category	Part Number	Product Description
	SW-2600-FDE-SKM-P	SW, FDE sec key management, 2600, -P
	SW-2600-SNAPSHOT-P	SW, Snapshot, 2600, -P
	SW-2600-VOL-COPY-P	SW, volume copy, 2600, -P
	SW-2600-REM-MIRR-P	SW, remote mirroring, 2600, -P
Software (add-on)	SW-5400-FDE-SKM	SW, FDE sec key management, 5400
	SW-5400-SNAPSHOT	SW, Snapshot, 5400
	SW-5400-VOL-COPY	SW, volume copy, 5400
	SW-5400-REM-MIRR	SW, remote mirroring, 5400
	SW-2600-FDE-SKM	SW, FDE sec key management, 2600
	SW-2600-SNAPSHOT	SW, Snapshot, 2600
	SW-2600-VOL-COPY	SW, volume copy, 2600
	SW-2600-REM-MIRR	SW, remote mirroring, 2600
Disk drives (packs)	E-X4021A-10-R6	Disk drives, 10x3TB, 7.2k, DE6600
	E-X4023A-10-R6	Disk drives, 10x2TB, 7.2k, DE6600
	E-X4022A-12-R6	Disk drives, 12x3TB, 7.2k, DE1600
	E-X4024A-12-R6	Disk drives, 12x2TB, 7.2k, DE1600
	E-X4027A-12-R6	Disk drives, 12x600GB, 3.5", 15k, DE1600
	E-X4025A-12-R6	Disk drives, 12x900GB, 2.5", 10k, DE5600
	E-X4026A-12-R6	Disk drives, 12x600GB, 2.5", 10k, DE5600
Single disk drives	E-X4021A-R6	Disk drive, 3TB, 7.2k, DE6600, QS
	E-X4023A-R6	Disk drive, 2TB, 7.2k, DE6600, QS
	E-X4022A-R6	Disk drive, 3TB, 7.2k, DE1600, QS
	E-X4024A-R6	Disk drive, 2TB, 7.2k, DE1600, QS

Category	Part Number	Product Description
	E-X4027A-R6	Disk drive, 600GB, 3.5", 15k, DE1600, QS
	E-X4025A-R6	Disk drive, 900GB, 2.5", 10k, DE5600, QS
	E-X4026A-R6	Disk drive, 600GB, 2.5", 10k, DE5600, QS
Expansion 10 packs	E-X4021A-10-QS-R6	Disk drives, 10x3TB, 7.2k, DE6600, QS
	E-X4023A-10-QS-R6	Disk drives, 10x2TB, 7.2k, DE6600, QS
Expansion 12 packs	E-X4022A-12-QS-R6	Disk drives, 12x3TB, 7.2k, DE1600, QS
	E-X4024A-12-QS-R6	Disk drives, 12x2TB, 7.2k, DE1600, QS
	E-X4027A-12-QS-R6	Disk drives, 12x600GB, 3.5", 15k, DE1600, QS
	E-X4025A-12-QS-R6	Disk drives, 12x900GB, 2.5", 10k, DE5600, QS
	E-X4026A-12-QS-R6	Disk drives, 12x600GB, 2.5", 10k, DE5600, QS
Spares/field-replaceable units	E-X4028A-R6	Solid-state drive, 800GB, 2.5", DE6600, QS
Disk drives single (field-replaceable unit only)	E-X4029A-R6	Solid-state drive, 200GB, 2.5", DE6600, QS
	E-X4030A-R6	Solid-state drive, 800GB, 2.5", DE5600, QS
	E-X4031A-R6	Solid-state drive, 200GB, 2.5", DE5600, QS
Miscellaneous hardware	X-48788-00-R6	Controller, E5400, 12GB, FC, no battery, SMID161
	X-24238-00-R6	Rail kit, DE1600, adjustable, 23.5"–32.5"
	X-41198-00-R6	Rail kit, DE6600, adjustable, 29.5"–35.75"
	X-48601-00-R6	Rail kit, DE6600, adjustable, 600–785mm
	X-48870-00-R6	PSU, 725W, AC, DE1600
	X-48564-00-R6	PSU, 1755W, AC, DE6600
	X-46381-00-R6	Battery, E2600
	X-48619-00-R6	Battery, E5400
	X-48565-00-R6	FAN, DE6600

Category	Part Number	Product Description
	X-48566-00-R6	Drawer, 12-drive, DE6600
	X-48567-00-R6	Bezel, front panel, DE6600
	X-24936-00-R6	Cable, mini-SAS, 2m, R6
	X-37953-00-R6	SFP, 8Gb, FC, E-Series
Power cords	X-50613-00-R6	Power cord, in-cabinet, 2m, C14-C19, 250V, DE6600
	X-52197-00-R6	Power cord, in-cabinet, 2m, C14-C13, E-Series
	X-33106-00-R6	Power cord, North America, 220V, E-Series
	X-33107-00-R6	Power cord, North America, 110V, E-Series
	X-33108-00-R6	Power cord, Europe, E-Series
	X-33109-00-R6	Power cord, Switzerland, E-Series
	X-33110-00-R6	Power cord, Italy, E-Series
	X-33111-00-R6	Power cord, UK and Ireland, E-Series
	X-33112-00-R6	Power cord, Denmark, E-Series
	X-33113-00-R6	Power cord, India, E-Series
	X-33115-00-R6	Power cord, Australia-New Zealand, E-Series
	X-33116-00-R6	Power cord, Israel, E-Series
	X-33117-00-R6	Power cord, China, E-Series
	X-41592-00-R6	Power cord, Taiwan, E-Series

3 Management of E-Series

3.1 E-Series SANtricity ES 10.80 Out-of-Band Management

Overview

SANtricity® ES is the GUI used to manage E-Series storage arrays. The application is based on the Java® framework and should be installed on a Windows® or Linux OS that does not participate in the data delivery workload. The installation package supports both 32-bit and 64-bit machines, and the installation procedure verifies that the application is being installed on the correct OS version. The management computer must have IP connectivity to each E-Series array controller to be managed. The SANtricity ES management client in the out-of-band configuration enables storage administrators to perform the following tasks:

- Commission new storage devices
- Set up network connections
- Provision storage and hosts

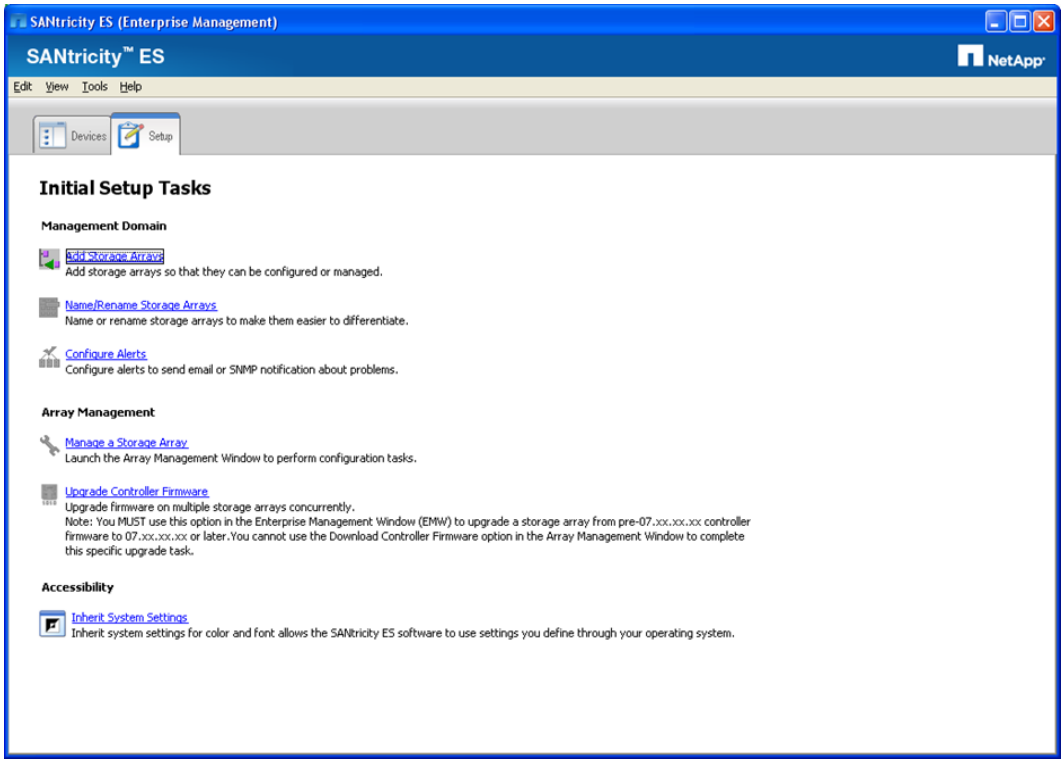
- Perform various maintenance activities to manage storage on E-Series storage arrays

When the SANtricity ES management client is installed on a desktop OS, the following limitations apply:

- Simultaneous user sessions are limited to eight sessions.
- Real-time system monitoring is not enabled.
- Desktop systems cannot run the host agent and send I/O to the E-Series storage array.

When the SANtricity ES management client is installed on a compute platform running a server OS, the full functionality is available; however, the number of simultaneous sessions is limited. Figure 8 shows the SANtricity ES Enterprise Management window.

Figure 8) SANtricity ES management client Enterprise Management window.



Guidelines

Follow these guidelines when using the SANtricity ES management client in the out-of-band configuration:

- Use out-of-band management when the storage administrator segregates management I/O from production I/O and during the initial commissioning operations that occur before hosts are connected to the array.
- The management server must access the storage array through an IP connection (DHCP or static) to the Ethernet management ports on the controller modules.
- Install the management application on a management node that does not participate in the data delivery workload.

Note: SANtricity ES out-of-band management is the preferred management method for E-Series arrays. However, in-band management is supported on a management server with FC or SAS connectivity to each array that will be managed. For more information about SANtricity ES in the in-band configuration, refer to the SANtricity ES 10.80 online help documentation.

4 Physical Infrastructure for E-Series

4.1 E-Series E5400 Hardware

Overview

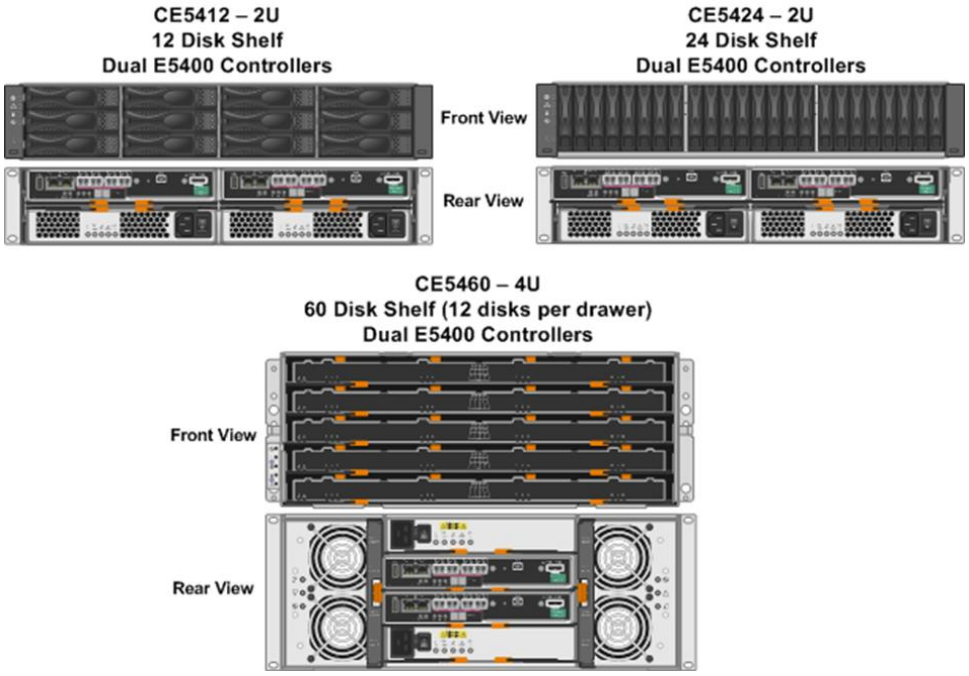
High-bandwidth applications and HPC platforms require high-performance, reliable, and scalable storage systems. The E5400-based storage system meets these requirements by supporting:

- Four 8Gb FC host interfaces per controller standard
- Multiple optional HICs, one per controller:
  - Four-port 8Gb FC
  - Two-port 40Gb IB
- 384 total disk drives per storage array
- Multiple RAID levels (0, 1, 10, 3, 5, and 6)
- A range of drive speeds and capacities
- Data assurance (T10-PI data integrity checking)
- Media parity check and correction capability
- Extensive event logging
- Recovery Guru onboard system diagnostics and recovery capability
- Hardware redundancy
- 6GB cache memory per controller (12GB optional) to maximize read/write performance
- NVSRAM and onboard USB drive to preserve the system configuration during power outages

As shown in Figure 9, the E5400 controller is available in three shelf packages (E5460, E5424, and E5412), each supporting dual controller canisters, power supplies, and fan units for hardware redundancy. The shelves are sized to support 60 disks, 24 disks, or 12 disks, respectively. Multiple disk expansion shelves (DE6600, DE5600, and DE1600) can be connected to the controller shelf to add additional storage capacity. For additional details, refer to the NetApp E5400 Storage System datasheet.

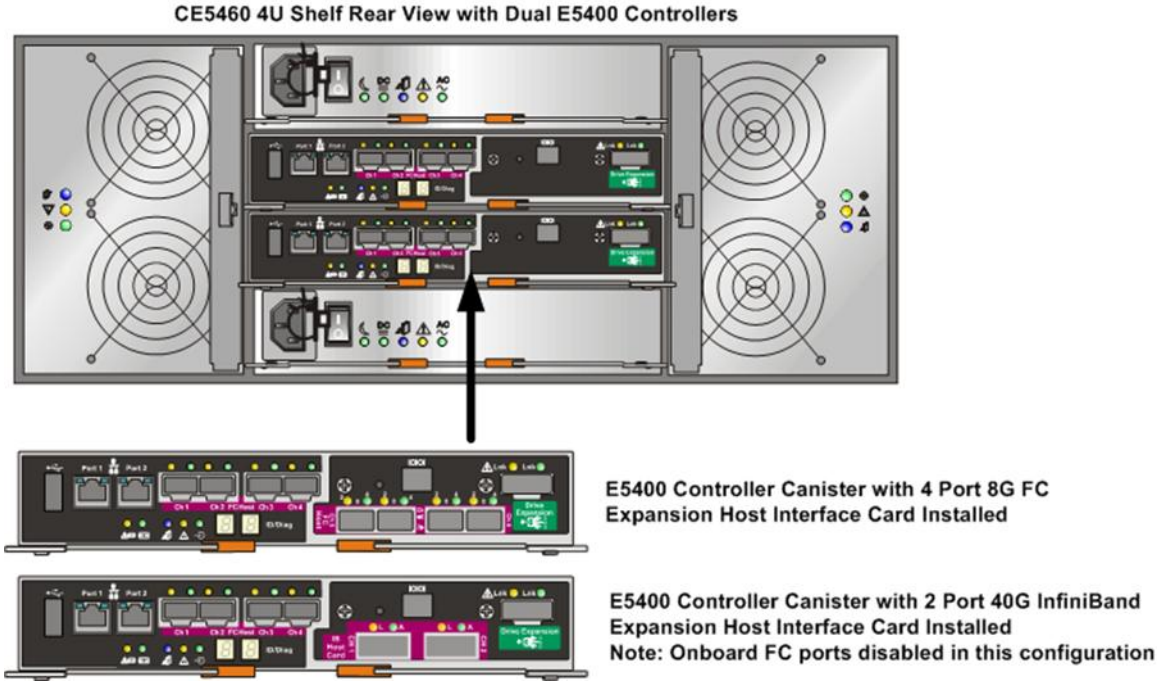
Note: The DE6600 60-disk E5400-based arrays should not exceed six total shelves counting the controller shelf, and the DE5600 and DE1600 shelf configurations should not exceed eight total shelves. Empty slots in any attached disk shelf are counted as drives when calculating the total drive count on an array.

Figure 9) E5400 shelf options.



By default, the E5400 controller canister has four 8Gb onboard FC ports for host-side communication channels, but it also supports channel expansion through add-on modules that add either four FC ports or two IB ports (onboard FC ports are disabled when the IB module is installed). Figure 10 shows the E5460 4U shelf with the available channel adapter modules.

Figure 10) E5460 controller shelf with optional host-side expansion ports.

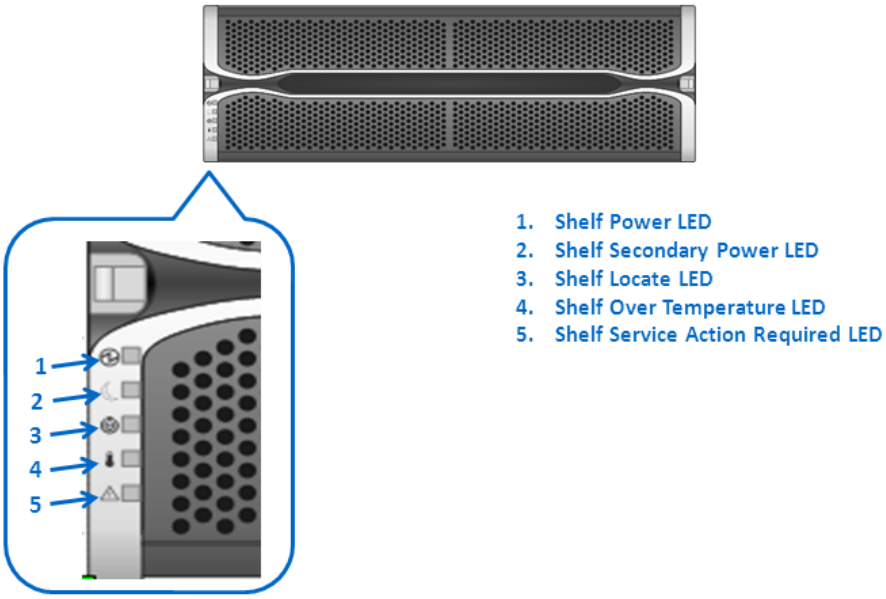


LED Status Indicators

Controller Drive Shelf LED Status Indicators

The E5400 controller shelf enclosure has several LEDs that indicate the overall status of the array, as shown in Figure 11.

Figure 11) Controller drive shelf status LEDs.



The shelf status LED layout is the same for all three packaging options (DE6600, DE5600, and DE1600). Table 10 lists the meanings of all the indicators.

Table 10) Controller drive shelf LED status definitions.

LED Name	Color	LED On	LED Off
Controller drive shelf power	Green	Power is present.	Normal status
Controller drive shelf secondary power	Green	Battery is fully charged. LED blinks when battery is charging.	Controller canister is operating without battery, or existing battery has failed.
Controller drive shelf locate	White	Identifies controller drive tray when SANtricity ES locate feature is activated.	Normal status
Controller drive shelf over temperature	Amber	The temperature of the controller drive tray has reached unsafe level.	Normal status
Controller drive shelf service action required	Amber	A component within the controller drive tray requires attention.	Normal status

Controller Base Features LED Status Indicators

The E5400 controller has several onboard LED status indicators, as shown in Figure 12. Most of the LEDs are lit when a fault condition exists; however, the battery charging and the cache-active LEDs are lit when the battery is fully charged and the cache is active. The seven-segment LEDs provide status codes for both normal operation and fault conditions, and the dot in the first seven-segment LED is the controller heartbeat indicator.

Figure 12) E5400 controller status indicator LEDs.

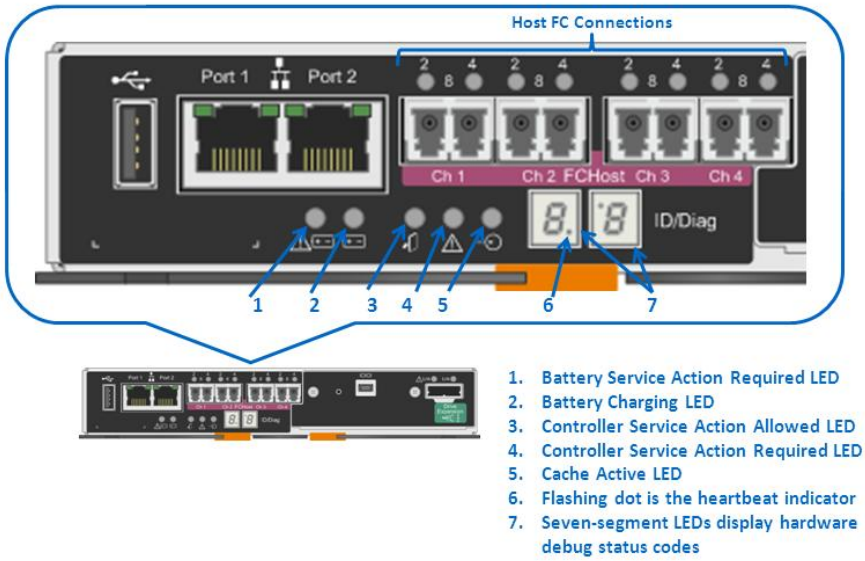


Table 11 provides additional controller status LED definitions.

Table 11) Controller base features LED status definitions.

LED Name	Color	LED On	LED Off
Battery service action required LED	Amber	Battery in controller canister has failed.	Normal status
Battery charging LED	Green	Battery is fully charged. LED blinks when battery is charging.	Controller canister is operating without battery, or existing battery has failed.
Controller service action allowed LED	Blue	Controller canister can be removed safely from controller drive tray.	Controller canister cannot be removed safely from controller drive tray.
Controller service action required LED	Amber	Some fault exists within controller canister.	Normal status
Cache active LED	Green	Cache is active. After AC power failure, this LED blinks while cache offload is in process.	Cache is inactive, or controller canister has been removed from controller drive tray.
Dot in lower-right corner of first seven-segment LED	Yellow (not amber)	Dot flashing indicates controller heartbeat is active	Dot not lit indicates controller heartbeat is not active (that is, controller is not in service).

LED Name	Color	LED On	LED Off
Two seven-segment LEDs	Yellow (not amber)	If controller status code = 99, then controller is in service. If controller status code does not = 99, then fault condition exists. Contact Technical Support for further assistance.	Controller is not powered on.

Note: The battery service action required LED indicates the battery timer has expired or the battery has failed the automatic battery test. This condition can seriously affect system write performance as the write cache feature is automatically disabled when the battery is not functioning normally.

Host-Side Ports LED Status Indicators

The host-side connection ports provide status LEDs to indicate the connection status for each link between the storage array and various host side hardware devices as shown in Figure 12. Table 12 and Table 13 provide the definitions for each LED.

Table 12) Ethernet management port status indicator definitions.

LED Name	Color	LED On	LED Off
Ethernet management port link rate LED (top-left corner of management port RJ-45 connectors)	Green	There is a 100BASE-T rate.	There is a 10BASE-T rate.
Ethernet management port connectors link active LED (top-right corner of management port RJ-45 connectors)	Green	Link is up (LED blinks when there is activity).	Link is not active.

Table 13) Host-side FC ports status indicator definitions.

FC Port LEDs (Link Active and Data Rate)	Color	LED On
Upper left = off, upper right = off	Green	Link not active.
Upper left = on, upper right = off	Green	Link active, data rate = 2Gb/sec
Upper left = off, upper right = on	Green	Link active, data rate = 4Gb/sec
Upper left = on, upper right = on	Green	Link active, data rate = 8Gb/sec

Drive-Side SAS Expansion Port

The E5400 controller canister is equipped with a 4-lane 6Gb/sec SAS expansion port used to connect additional disk shelves to the E5400 controller shelf. Figure 13 shows a close-up of the SAS expansion port LEDs.

Figure 13) E5400 drive expansion port status indicator LEDs.

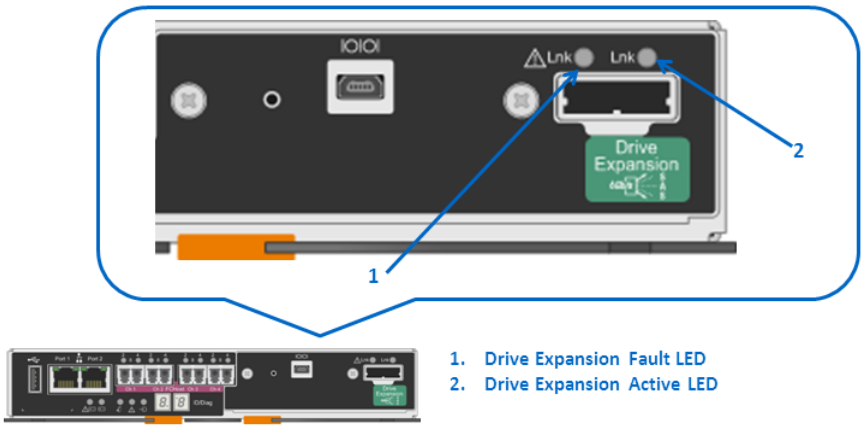


Table 14 provides the definitions for each drive-side LED.

Table 14) Drive-side SAS ports status indicator definitions.

LED Name	Color	LED On	LED Off
Drive expansion link fault	Amber	At least one of the four PHYs in out port is working, but another PHY cannot establish same link to expansion out connector.	Normal status
Drive expansion link active	Green	At least one of four PHYs in out port is working, and link exists to device connected to expansion out connector.	Link error has occurred.

For additional details on the E5400 controller and related hardware, refer to the NetApp E-Series Storage Systems CE5400 Controller-Drive Tray Installation Guide.

Guidelines

Consider the following guidelines when implementing the E5400 storage system:

- Determine the level of performance required by the compute platforms to support the given applications.
- Determine the amount of storage capacity required (include the number of disks required for hot spares).
- Choose the disk types based on performance and capacity requirements.
- Determine the power and network connectivity requirements.
- Plan RAID levels to achieve the level of reliability and read/write performance required.
- Determine which hosts will be connected to the storage system and plan the configuration of the storage system ports to maximize throughput.
- Plan to install and configure host multipath software to achieve host-side channel redundancy.
- Plan for management access to the storage platform by using either the in-band management or the out-of-band management methodology (out of band is most commonly used).

- Use the SANtricity ES client to connect to the storage system and to implement the planned configuration.
- Always save the system configuration and profile after configuration or provisioning changes so that in case of a catastrophic system fault the system can be fully recovered.

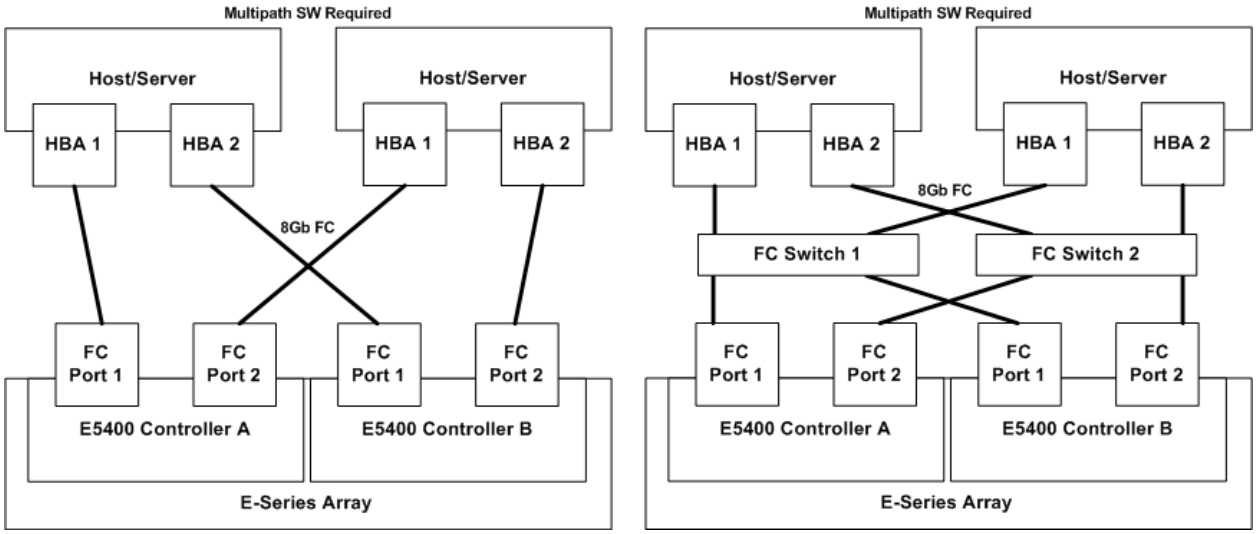
SANtricity ES is the GUI management interface for E-Series arrays. It is based on the Java framework and can be installed on Windows or Linux OSs. The management application should be installed on a management node that does not participate in production data delivery. The software is available in 32-bit and 64-bit versions, and the install process detects if the installation of the package is performed on the wrong OS version.

The SANtricity ES client software can be installed on Windows or Linux OS for out-of-band management of the storage array. In this configuration, the host agent functionality for in-band management does not function, and the number of client connections is limited to eight. To manage the storage arrays by using in-band connections, the management client must be running a server OS and have FC connectivity to all arrays. In this configuration, the eight-session maximum does not apply.

Additional Information

For host-side FC and IB connections, the hosts can be connected either directly to the storage controller or through a switch that allows multiple hosts to share the paths, as shown in Figure 14. Both configurations require multipath software on the host for link management.

Figure 14) Host connection examples.



4.2 E-Series E2600 Hardware

Overview

High-bandwidth applications and HPC platforms require high-performance, reliable, and scalable storage systems. The E2600-based storage system meets these requirements by supporting:

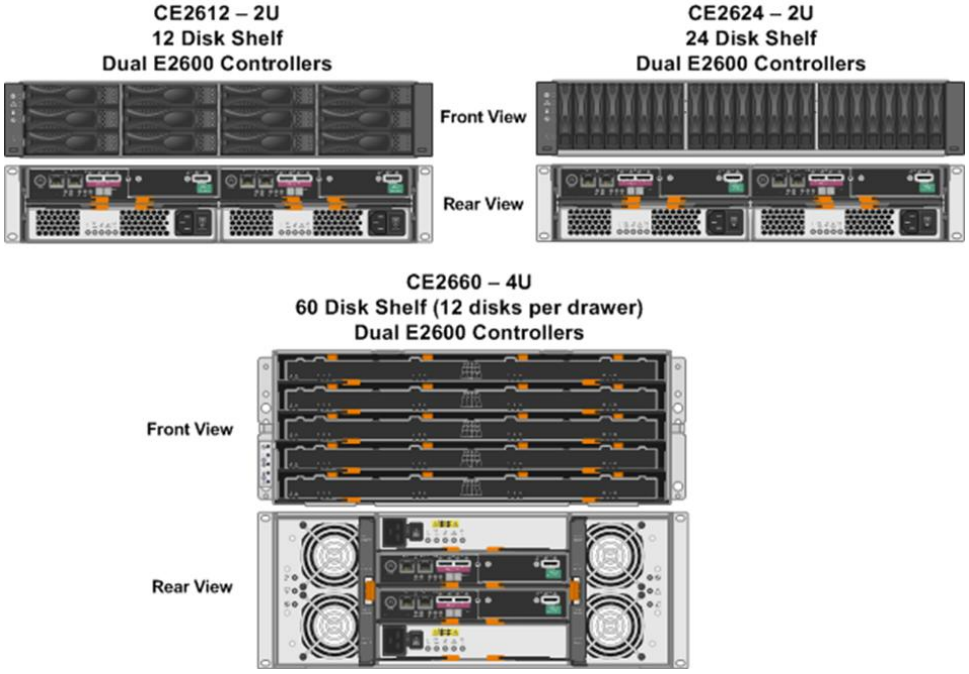
- Two 4-lane 6Gb SAS host interface ports per controller standard
- Multiple optional HICs, one per controller:
  - Four-port 8Gb FC
  - Four-port 1Gb iSCSI
  - Two-port 10Gb iSCSI

  - Two-port 6Gb SAS
- 192 total disk drives per storage array
- Multiple RAID levels (0, 1, 10, 3, 5, and 6)
- A range of drive speeds and capacities
- Data assurance (T10-PI data integrity checking)
- Media parity check and correction capability
- Extensive event logging
- Recovery Guru onboard system diagnostics and recovery capability
- Hardware redundancy
- 1GB cache memory per controller (2GB optional) to maximize read/write performance
- NVSRAM and onboard USB drive to preserve the system configuration during power outages

As shown in Figure 15, the E2600 controller is supported in three shelf packages (E2660, E2624, and E2612), each supporting dual controller canisters, power supplies, and fan units for hardware redundancy. The shelves are sized to support 60 disks, 24 disks, or 12 disks, respectively. Multiple disk expansion shelves (DE6600, DE5600, and DE1600) can be connected to the controller shelf to add additional storage capacity. For additional details, refer to the NetApp E2600 Storage System datasheet.

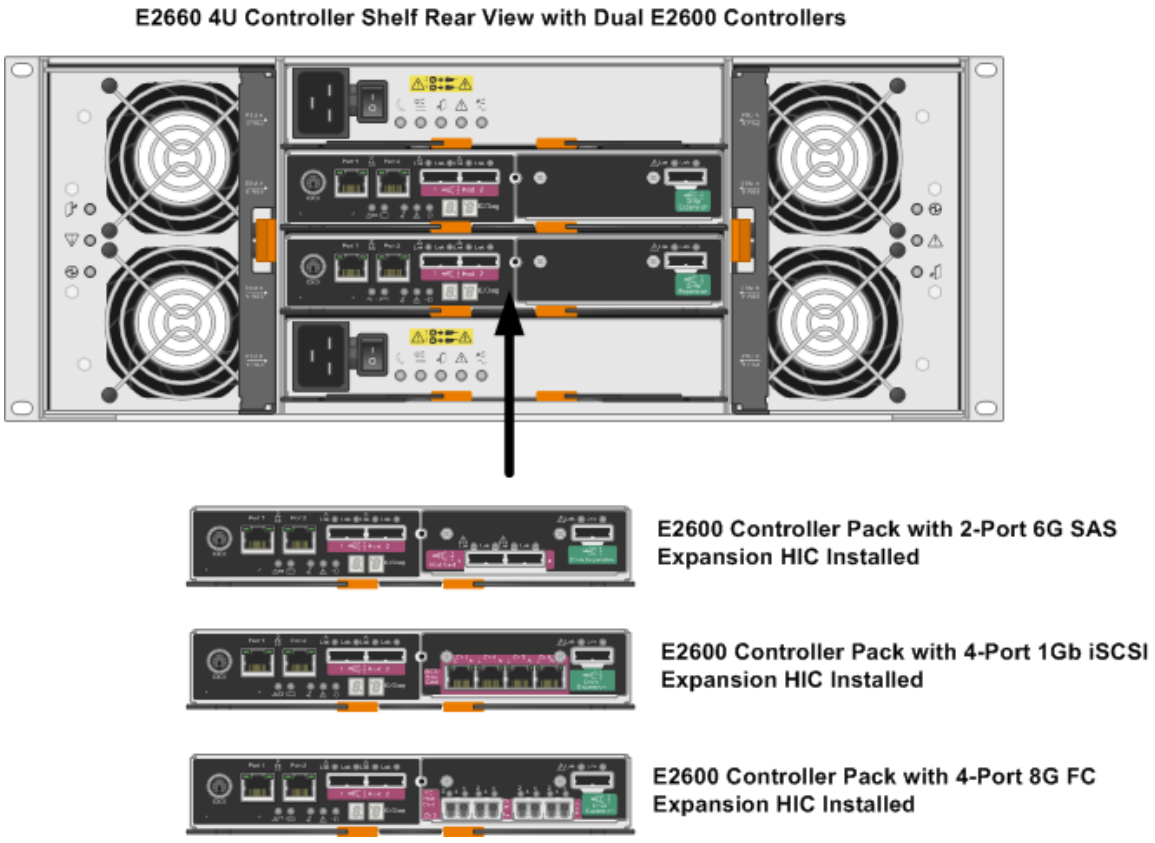
Note: Empty slots in any attached disk shelf are counted as drives when calculating the total drive count on an array.

Figure 15) E2600 shelf options.



By default, the E2600 controller canister has two onboard 4-lane 6Gb/sec SAS ports for host-side communication channels, but it also supports host-side channel expansion through add-on modules that add either four FC ports and four 1Gb iSCSI ports, two 10Gb iSCSI ports, or 2 additional 4-lane 6Gb SAS ports. Figure 16 shows the E2660 4U shelf with the available controller configurations.

Figure 16) E2660 controller with optional host-side expansion ports.

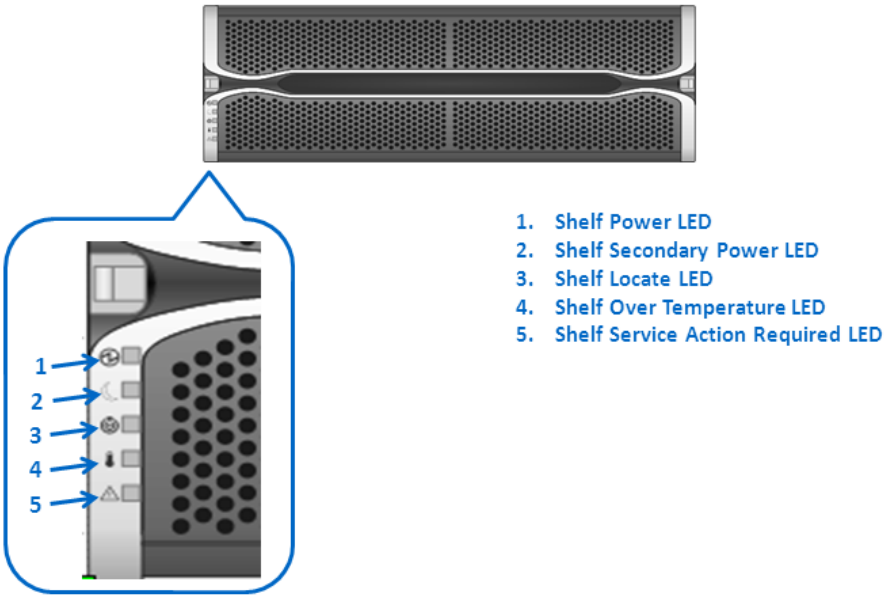


LED Status Indicators

Controller Drive Shelf LED Status Indicators

The E2600 controller shelf enclosure has several LEDs that indicate the overall status of the array, as shown in Figure 17.

Figure 17) Controller drive shelf status LEDs.



The shelf status LED layout is the same for all three packaging options (DE6600, DE5600, and DE1600). Table 15 lists the meanings of all indicators.

Table 15) Controller disk shelf LED status definitions.

LED Name	Color	LED On	LED Off
Controller drive shelf power	Green	Power is present.	Normal status
Controller drive shelf secondary power	Green	Battery is fully charged. LED blinks when battery is charging.	Controller canister is operating without battery, or existing battery has failed.
Controller drive shelf locate	White	Identifies controller drive tray when SANtricity ES locate feature is activated.	Normal status
Controller drive shelf over temperature	Amber	Temperature of controller drive tray has reached unsafe level.	Normal status
Controller drive shelf service action required	Amber	Component within controller drive tray requires attention.	Normal status

Controller Base Features LED Status Indicators

The E2600 controller has several onboard LED status indicators, as shown in Figure 18. Most of the LEDs are lit when a fault condition exists; however, the battery-charging and the cache-active LEDs are lit when the battery is fully charged and the cache is active. The seven-segment LEDs provide status codes for both normal operation and fault conditions, and the dot in the first seven-segment LED is the controller heartbeat indicator.

Figure 18) E2600 controller status indicator LEDs.

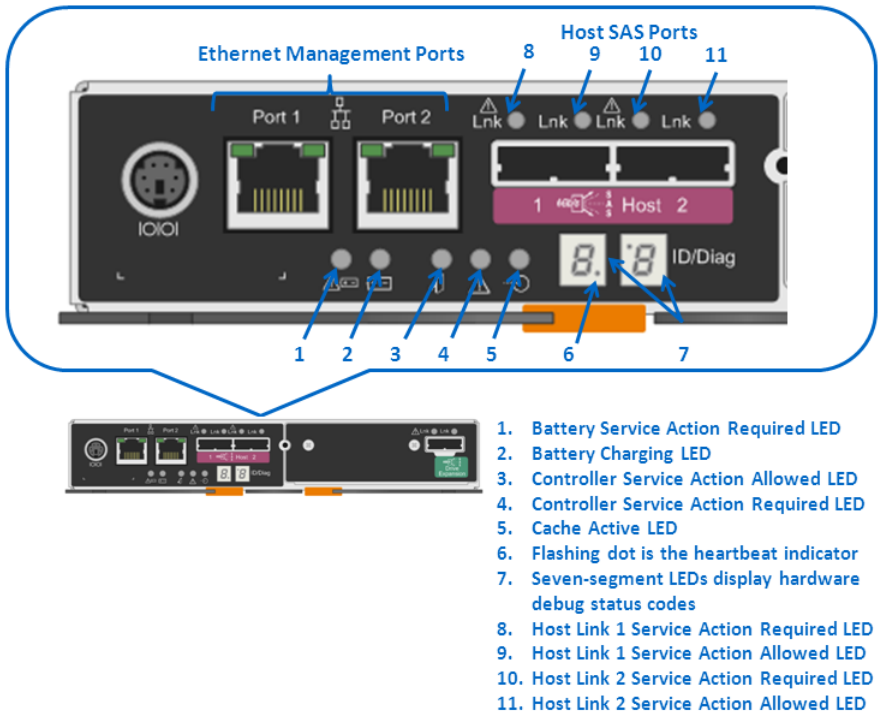


Table 16 provides additional controller status LED definitions.

Table 16) Controller base features LED status definitions.

LED Name	Color	LED On	LED Off
Battery service action required LED	Amber	Battery in controller canister has failed.	Normal status
Battery charging LED	Green	Battery is fully charged. LED blinks when battery is charging.	Controller canister is operating without battery, or existing battery has failed.
Controller service action allowed LED	Blue	Controller canister can be removed safely from controller drive tray.	Controller canister cannot be removed safely from controller drive tray.
Controller service action required LED	Amber	Some fault exists within controller canister.	Normal status
Cache active LED	Green	Cache is active. After AC power failure, this LED blinks while cache offload is in process.	Cache is inactive, or controller canister has been removed from controller drive tray.
Dot in lower-right corner of first seven-segment LED	Yellow (not amber)	Flashing dot indicates controller heartbeat is active.	Dot not lit indicates controller heartbeat is not active (that is, controller is not in service).

LED Name	Color	LED On	LED Off
Two seven-segment LEDs	Yellow (not amber)	If controller status code = 99, then controller is in service. If controller status code does not = 99, then fault condition exists. Contact Technical Support for further assistance.	Controller is not powered on.

Note: The battery service action required LED indicates that the battery timer has expired or the battery has failed the automatic battery test. This condition can seriously affect the system write performance because the write cache feature is automatically disabled when the battery is not functioning normally.

Host-Side Ports LED Status Indicators

The host-side connection ports provide status LEDs to indicate the connection status for each link between the storage array and various host-side hardware devices as shown in Figure 18. Table 17 and Table 18 provide the definitions for each LED.

Table 17) Ethernet management port status indicator definitions.

LED Name	Color	LED On	LED Off
Ethernet management port link rate LED (top-left corner of management port RJ-45 connectors)	Green	There is a 100BASE-T rate.	There is a 10BASE-T rate.
Ethernet management port connectors link active LED (top-right corner of management port RJ-45 connectors)	Green	Link is up (LED blinks when there is activity).	Link is not active.

Table 18) Host-side SAS ports status indicator definitions.

LED Name	Color	LED On	LED Off
Host link 1 service action required LED	Amber	At least one of four PHYs is working, but another PHY cannot establish same link to device connected to host in port connector.	No link error has occurred.
Host link 1 service action allowed LED	Green	At least one of four PHYs in host in port is working, and link exists to device connected to in port connector.	No link error has occurred.
Host link 2 service action required LED	Amber	At least one of four PHYs is working, but another PHY cannot establish same link to device connected to host in port connector.	No link error has occurred.

LED Name	Color	LED On	LED Off
Host link 2 service action allowed LED	Green	At least one of four PHYs in host in port is working, and link exists to device connected to in port connector.	No link error has occurred.

Drive-Side SAS Expansion Port

The E2600 controller canister is equipped with a SAS expansion port used to connect additional disk shelves to the E2600 controller shelf. Figure 19 shows a close-up of the SAS expansion port LEDs.

Figure 19) E2600 drive expansion port status indicator LEDs.

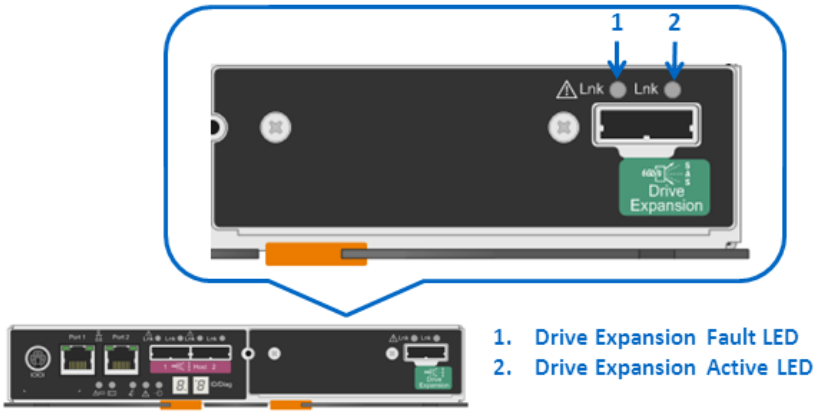


Table 19 provides the definitions for each drive-side LED.

Table 19) Drive-side SAS ports status indicator definitions.

LED Name	Color	LED On	LED Off
Drive expansion link fault	Amber	At least one of four PHYs in out port is working, but another PHY cannot establish same link to expansion out connector.	Normal status
Drive expansion link active	Green	At least one of four PHYs in out port is working, and link exists to device connected to expansion out connector.	Link error has occurred.

For additional details on the E2600 controller and related hardware, refer to the NetApp E-Series Storage Systems CE2600-60 Controller-Drive Tray Installation Guide.

Guidelines

Consider the following guidelines when implementing the E2600 storage system:

- Determine the level of performance required by the compute platforms to support the given applications.

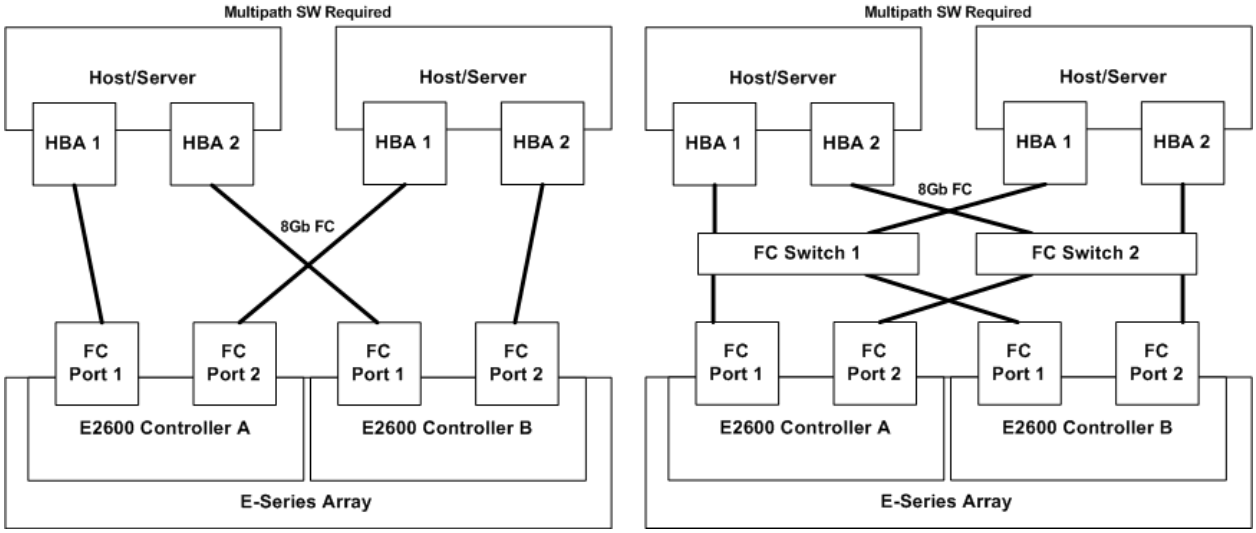
- Determine the amount of storage capacity required (include the number of disks required for hot spares).
- Choose the disk types based on performance and capacity requirements.
- Determine the power and network connectivity requirements.
- Plan RAID levels to achieve the level of reliability and read/write performance required.
- Determine which hosts will be connected to the storage system and plan the configuration of the storage system ports to maximize throughput.
- Plan to install and configure host multipath software to achieve host-side channel redundancy.
- Plan for management access to the storage platform by using either the in-band management or the out-of-band management methodology (out of band is most commonly used).
- Use the SANtricity ES client to connect to the storage system and to implement the planned configuration.
- Always save the system configuration and profile after configuration or provisioning changes so that in case of a catastrophic system fault the system can be fully recovered.

SANtricity ES is the GUI management interface for E-Series arrays and operates in an in-band or in an out-of-band mode, but the management application should be installed on a management node that does not participate in production data delivery. The GUI is based on the Java framework and can be installed on Windows or Linux OSs. The software is available in 32-bit and 64-bit versions. The install process detects the OS and hardware version and installs the correct version for that platform. To manage the storage arrays by using in-band connections, the management client must be running a server OS and have FC connectivity to all arrays.

Additional Information

For host-side FC and iSCSI connections, the hosts can be connected either directly to the storage controller or through a switch that allows multiple hosts to share the paths, as shown in Figure 20. Both configurations require multipath software for link management.

Figure 20) Host connection examples.



SAS ports are usually cabled directly to local servers. In this configuration, make sure all host servers have a path to both controllers in E-Series arrays and install the appropriate multipath software for the server OS type.

4.3 E-Series Disk Expansion Shelves

Overview

E-Series arrays support storage capacity growth beyond the disk slots in the controller shelf by adding disk expansion shelves to new or existing E5400- and E2600-based storage arrays. The additional DE6600 (60-disk), DE5600 (24-disk), or DE1600 (12-disk) shelf enclosures have environmental services monitor (ESM) canisters installed instead of controller canisters. Figure 21 shows the ESM canister.

Figure 21) ESM canister.

E-Series disk expansion shelves can be added in combinations of 4U and 2U packages to achieve specific performance and capacity requirements. The typical configurations for each shelf type shown in Figure 22, Figure 23, and Figure 24 represent best practice cabling topology to maximize system resiliency against shelf hardware fault scenarios.

Note: The ESM canister has two SAS input ports and one output port for intershelf cabling. Never connect to both input ports on an ESM canister. Only one of the input ports can be used.

Figure 22) Maximum capacity E-Series array configuration using DE6600 shelves.

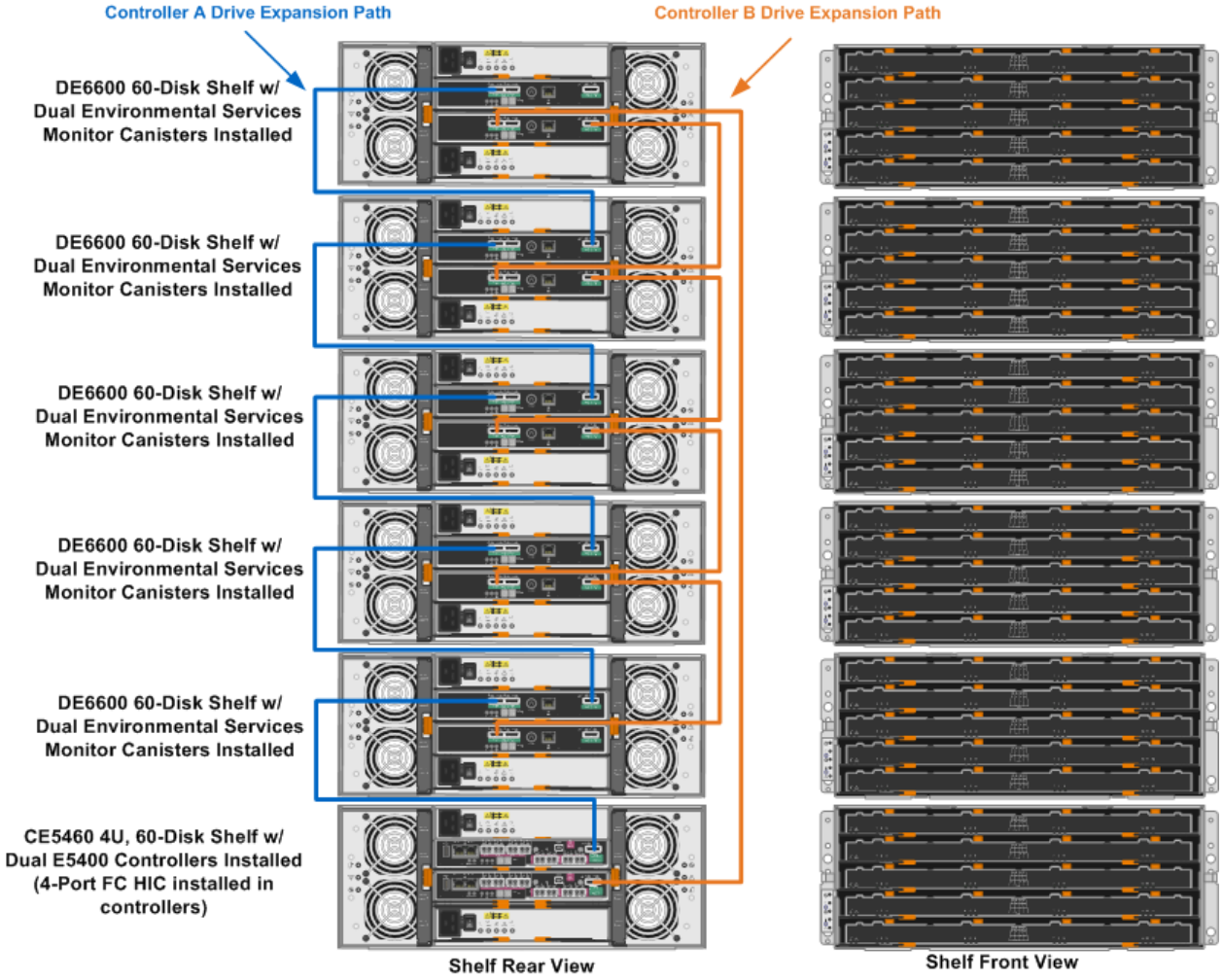


Figure 23) Typical E-Series array configuration using DE5600 shelves.

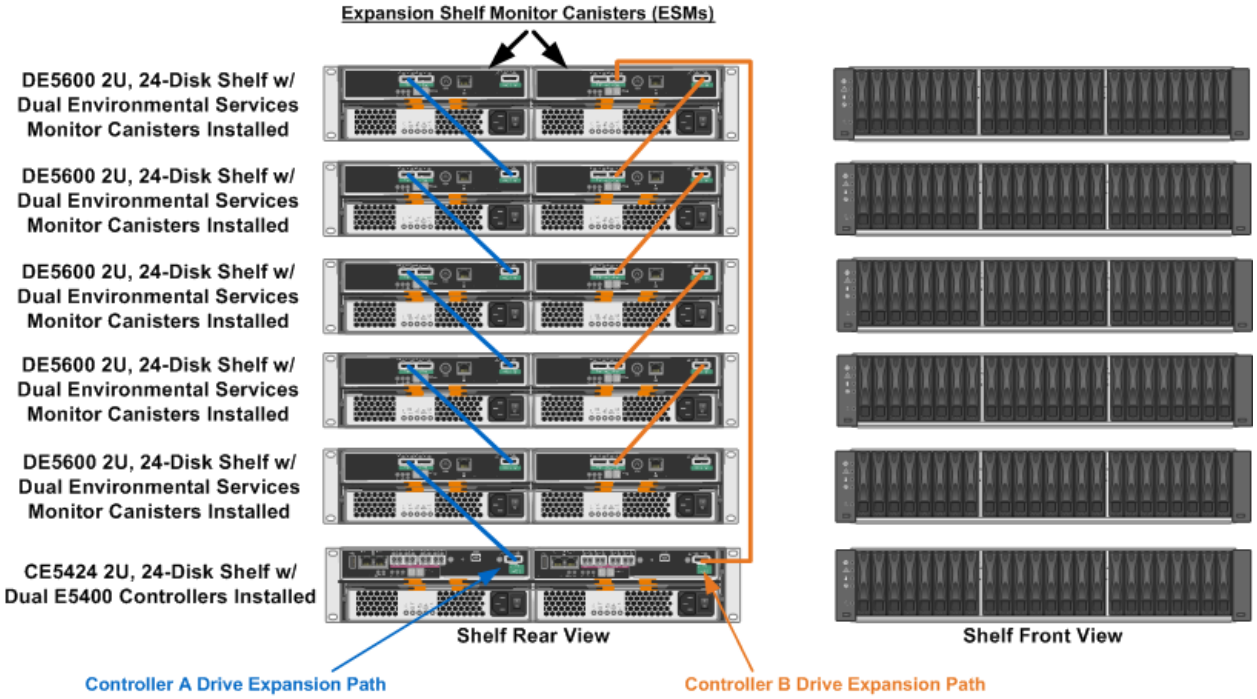
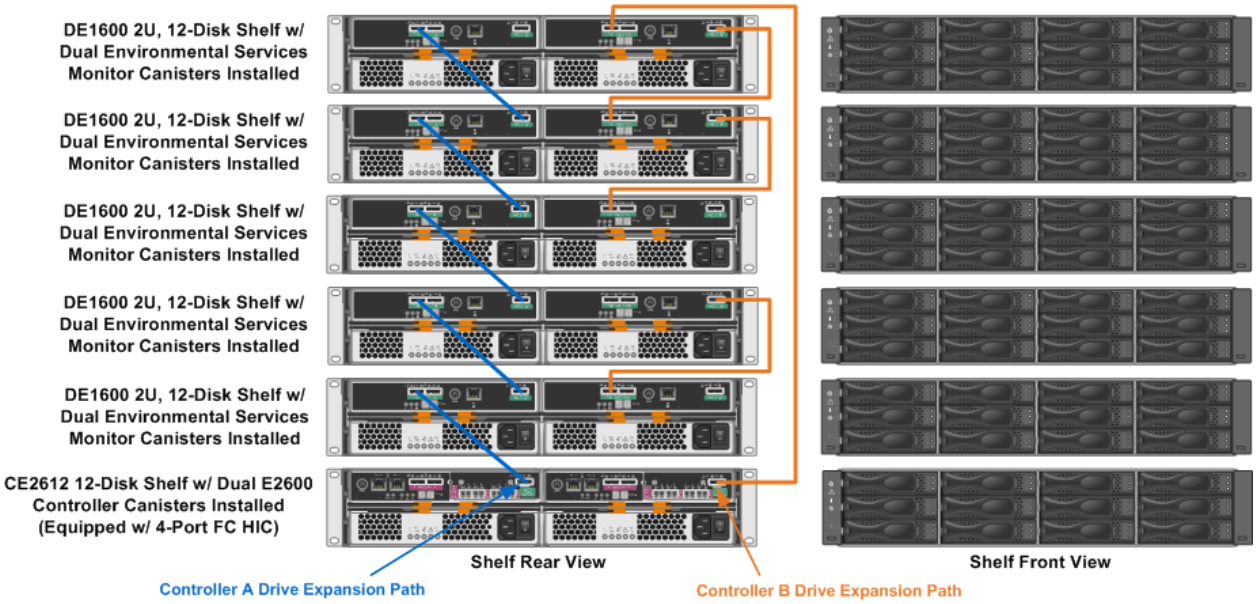


Figure 24) Typical E-Series array configuration using DE1600 shelves.



Note: NetApp does not recommend mixing shelf models in the same array because of the differences in the drives and the supported features.

The disk shelves must be installed within six feet of the array controller shelf to allow SAS cables to reach from the SAS expansion ports on the controller canisters to the ESM canisters installed in a disk shelf or from the SAS expansion port on one expansion shelf to the ESM SAS input port on a successive disk expansion shelf.

For additional details on how to install the E5400-based array configurations, refer to the NetApp E-Series Storage Systems CE5400 Controller-Drive Tray Installation Guide. For additional details on how to install the E2600-based storage array, refer to the NetApp E-Series Storage Systems CE2600-60 Controller-Drive Tray Installation Guide.

The most common configuration for an array is to stack the disk expansion shelves in the same physical rack that contains the controller shelf. However, care should be taken to ascertain that floor loading limitations are not exceeded. Whenever possible, the 60-disk chassis should be placed in the lower portion of the racks. Appropriate lifting equipment should be used to mount the chassis because a single shelf with the disks installed can exceed 200 pounds.

The E-Series E5400-based storage array can support a maximum of 384 individual drives, while the E2600-based storage array can support up to 192 individual drives. However, depending on shelf model and disk selection, additional boundaries to drive count must be considered. For more information on supported configurations, refer to the NetApp E5400 Storage System datasheet or to the NetApp E2600 Storage System datasheet.

Note: Empty disk slots in any connected disk shelf still count as a disk for planning the total disk count for a single storage array.

When initially powering on an E-Series array with disk expansion shelves, power on the disk shelves first, and wait one to two minutes before powering on the controller shelf. To add a disk expansion shelf to an existing E-Series array, follow the specific installation steps. For more information and assistance with adding a disk expansion shelf to an existing production E-Series array, contact NetApp Global Services.

Guidelines

Follow these best practices when installing and configuring the E5400 storage system:

- Do not mix different drive speeds in the same drive shelf drawer.
- Do not use both SAS input ports on an ESM at the same time.
- Install the CE5460 shelves at the bottom of the racks to prevent the rack from becoming top-heavy.
- Use lifting equipment when mounting controller and disk shelves because these shelves can exceed 200 pounds when they are fully loaded with disks.
- Replace failed disks with disks that match the failed disk:
  - SSDs must be replaced by other SSDs.
  - Encryption-capable drives must be replaced by other encryption-capable drives.
  - Drives supporting T10PI data assurance must be replaced by T10PI-capable drives.
- Route disk channel cables to avoid single points of failure.

Note: For additional cabling guidance for specific storage array models, refer to the NetApp E-Series Storage Systems Hardware Cabling Guide on the NetApp Support site.

5 Storage for E-Series

5.1 E-Series OST Configuration for Lustre File Systems

Overview

NetApp E-Series solutions that use the Lustre file system support several host configurations that enable OSS-to-OST workflows.

E5460 Overview

Utilize the full capacity of the E5460/DE6600 array for the Lustre file system by using SANtricity ES to create a base layout containing six RAID 6 (8+2) volume groups per 60-drive array. Use the entire capacity of the group to create one volume in each group and map each LUN to the group containing the OSSs.

Note: RAID 6 provides protection from dual-disk failure scenarios. However, if all available disks in the array are allocated in volume groups, NetApp recommends that any disk failure scenario be treated as a critical service event and that faulty disks be replaced as soon as possible after failure. Customers should purchase and store spare disks for easy access in case a disk-related service event occurs.

Although the base architecture associates a single OSS with a single NetApp E-Series E5460 storage array with six LUNs or OSTs, different HA configurations are commonly used in Lustre implementations. Multiple configurations can be valid; therefore, plan the configuration carefully to maximize throughput and load balancing from the perspective of the E-Series controllers. This can be accomplished by balancing the E-Series controller LUN (OST) ownership evenly across the LUNs that are mapped to any one OSS. However, this balanced approach is not always physically possible with the base layout (six RAID 6 (8+2) volume groups with one volume per group).

For example, in a configuration with two OSSs connected to one E-Series E5460, each OSS has active access to three OSTs (LUNs), two OSTs owned by one E-Series array controller, and one OST owned by the other controller. In this configuration, workflows must be carefully considered and monitored to make sure that the E-Series controllers do not become bottlenecks that limit performance.

For additional information about file system planning, refer to NetApp Technical Report 4006i, “NetApp High Performance Computing Solution for Lustre: Sizing Guide,” available on Field Portal.

When configuring the array, use a standard naming convention that associates the volume group and volume names to the OSS. This naming convention is helpful if troubleshooting path faults becomes necessary. Table 20 contains suggested naming conventions. These names should be determined during the planning phase of the project and used while configuring the storage.

Table 20) OST HA storage configuration.

Primary E5460	Secondary E5460
E5460-1_volgrp01 E5460-1_volgrp01_vol01_OSS ID LUN 1	E5460-2_volgrp07 E5460-2_volgrp07_vol01_OSS ID LUN 7
E5460-1_volgrp02 E5460-1_volgrp02_vol01_OSS ID LUN 2	E5460-2_volgrp08 E5460-2_volgrp08_vol01_OSS ID LUN 8
E5460-1_volgrp03 E5460-1_volgrp03_vol01_OSS ID LUN 3	E5460-2_volgrp09 E5460-2_volgrp09_vol01_OSS ID LUN 9
E5460-1_volgrp04 E5460-1_volgrp04_vol01_OSS ID LUN 4	E5460-2_volgrp10 E5460-2_volgrp10_vol01_OSS ID LUN 10
E5460-1_volgrp05 E5460-1_volgrp05_vol01_OSS ID LUN 5	E5460-2_volgrp11 E5460-2_volgrp11_vol01_OSS ID LUN 11

Primary E5460	Secondary E5460
E5460-1_volgrp06 E5460-1_volgrp06_vol01_OSS ID LUN 6	E5460-2_volgrp12 E5460-2_volgrp12_vol01_OSS ID LUN 12

Note: Delete the access LUN (LUN 7 is the default) if in-band management is not used. The access LUN should be deleted when the array is first discovered by SANtricity ES. The default for this implementation is out-of-band management through SANtricity ES.

E5424 Overview

Utilize 20 out of every 24 drives of the E5424/DE5600 array for the Lustre file system by using SANtricity ES to create a base layout containing two RAID 6 (8+2) volume groups per 24-drive array and assign the remaining drives as hot spares. Use the entire capacity of the group to create one volume in each group and map each LUN to the group containing the OSSs.

Although the base architecture associates a single OSS to a single NetApp E-Series E5424 storage array with two LUNs or OSTs, different HA configurations are commonly used in Lustre implementations. Multiple configurations can be valid; therefore, carefully plan the configuration to maximize throughput and load balancing from the perspective of the E-Series controllers. This can be accomplished by balancing the E-Series controller LUN (OST) ownership evenly across the LUNs that are mapped to any one OSS. Adding LUNs in pairs and mapping the added LUNs to alternate controllers makes the array physically balanced across controllers. Furthermore, workflows must also be carefully considered and monitored to make sure that the E-Series controllers do not become bottlenecks that limit performance.

For additional information about file system planning, refer to NetApp Technical Report 4006i, “NetApp High Performance Computing Solution for Lustre Sizing Guide,” available on Field Portal.

When configuring the array, use a standard naming convention that associates the volume group and volume names with the OSS. This naming convention is helpful if it becomes necessary to troubleshoot path faults. Table 20 contains suggested naming conventions. These names should be determined during the planning phase of the project and used while configuring the storage.

E5460/E5424 Storage Guidelines

Follow these guidelines when determining the appropriate OST configuration for E-Series solutions that use the Lustre file system:

- Determine whether the disk initialization time for volume group creation is a concern.
- Use SANtricity ES to create six RAID 6 (8+2) volume groups with one volume per group. When creating new groups, use the entire capacity of the volume group.
- Name the volumes to reflect the path between the OSS and the primary OST.
- Monitor the system by using SANtricity ES and set up alerts to make sure fault conditions are detected and addressed in a timely manner.
- Purchase and store spare disks (especially when using the E5460 array), so they are readily available to replace a faulty disk.
- If hot spare disks are not assigned on the array, treat any disk-related fault condition as a critical service event and replace the faulty disk as soon as possible.
- Map all OSTs (LUNs) to the default group for OSS access and sharing in HA configurations.
- Assign OST array controller ownership in a manner that is as balanced as possible. Monitor performance over time to make sure that the assignment achieves the best performance for the environment.

5.2 E-Series MDT Configuration for Lustre File Systems

Overview

When NetApp E-Series E2624 arrays are used to satisfy the metadata storage requirement for a Lustre file system, one RAID 10 volume group is sized according to the sizing guidance provided for the NetApp HPC Solution for Lustre in NetApp Technical Report 4006i, “NetApp High Performance Computing Solution for Lustre: Sizing Guide,” available on Field Portal. To meet the MDS storage requirement, use SANtricity ES to create one RAID 10 volume group and one MDT LUN with the default segment size of 128KB, using the entire capacity of the volume group.

Note: RAID 10 provides protection from disk failure scenarios by writing a duplicate copy of the data on each disk to a second disk, thus forming mirrored pairs of disks. In addition to this level of protection, NetApp recommends provisioning at least one spare disk to enable the automated disk rebuild process to function in the event of a failed disk.

If all available disks in the array are allocated in volume groups, NetApp recommends that any disk failure scenario be treated as a critical service event and that faulty disks be replaced as soon as possible after failure. Customers should purchase and store spare disks for easy access in case a disk-related service event occurs.

Map the MDT to the E2624 MDS host group to enable shared access from the MDSs. Cluster-management software on the servers is used to monitor server heartbeats and to determine which MDS is active and which is passive.

Guidelines

Follow these guidelines when determining the appropriate MDT configuration for E-Series solutions that use the Lustre file system:

- Confirm that each MDS is associated with the MDS host group on the E-Series array.
- Use the SANtricity ES automatic configuration feature to create a volume group and a volume (MDT). To determine the growth estimates for the specific Lustre environment, refer to NetApp Technical Report 4006i, “NetApp High Performance Computing Solution for Lustre: Sizing Guide,” available on Field Portal.
- Map the MDT to the E2624 MDS host group to enable MDS shared access to the MDT LUN.
- Monitor the system by using SANtricity ES and set up alerts so that fault conditions are detected and addressed in a timely manner.
- Allocate at least one spare disk per 24-disk array or purchase spare disks to have readily available in case a disk failure occurs.
- This configuration uses RAID 10 disk protection. Therefore, the volume group is protected from multiple disk failure scenarios. However, for scenarios in which both disks in a mirrored pair fail, the associated data is lost. NetApp recommends treating any disk failure scenario as a critical service event.

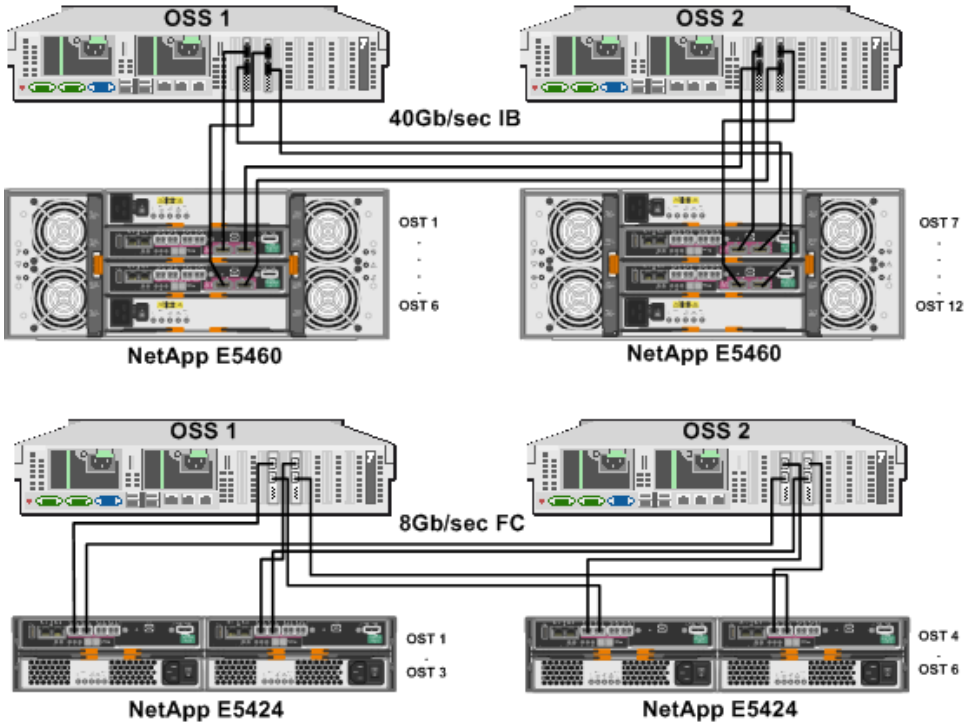
6 Operating Systems Connecting to E-Series

6.1 E-Series OSS Configuration for Lustre File Systems

Overview

NetApp E-Series solutions that use the Lustre file system support several host configurations that enable OSS-to-OST workflows. Although the base architecture associates a single OSS to a single NetApp E-Series E5424 or E5460 storage array, NetApp recommends an HA configuration with two hosts connected to two E-Series arrays in a cross-array configuration, as shown in Figure 25.

Figure 25) Lustre HA OSS-to-storage architectures.

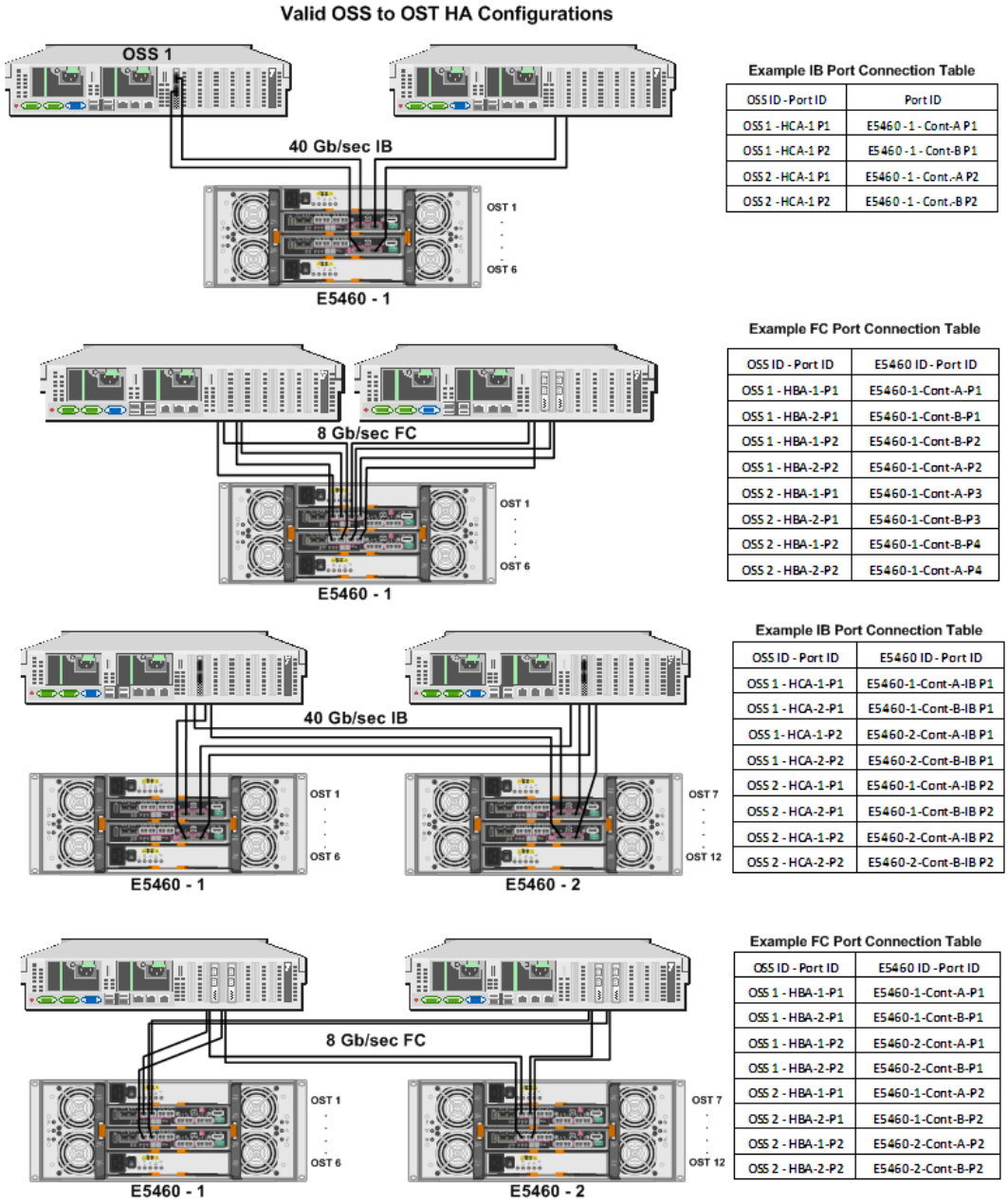


The example in Figure 25 shows two E5460 arrays with two OSSs in a direct connection IB configuration. The E5424 configuration offers a storage alternative to the E5460, but both options are configured in a similar manner. To build the OSS storage configuration on the E5424 or E5460 E-Series arrays, use SANtricity ES to create a single host (OSS), called primary host, and secondary hosts in an HA configuration. During the primary host creation procedure, indicate that the host will share storage with other hosts (specifically with the OSS HA pair). Once the links are identified, the Host Creation wizard prompts for a host group name. Use a descriptive name that facilitates the management of the storage environment over time. Map all of the available LUNs on the array to the host group.

When adding a secondary OSS host to the array, indicate again that the host will share storage with other hosts. The wizard prompts for a host group name, but instead of creating a second host group, select the host group name that was created with the initial OSS host. This configuration enables the host-side software to provide active-passive management of the hosts and all available storage on the array.

When the OSS has multiple FC or IB links to the E-Series array, install the SANtricity ES RDAC multipath driver to manage the multiple paths. Figure 26 shows additional typical configurations for Lustre OSS HA architectures.

Figure 26) Common OSS-to-OST HA configurations.



Example IB Port Connection Table

OSS ID - Port ID	Port ID
OSS 1 - HCA-1 P1	E5460 -1 - Cont-A P1
OSS 1 - HCA-1 P2	E5460 -1 - Cont-B P1
OSS 2 - HCA-1 P1	E5460 -1 - Cont.-A P2
OSS 2 - HCA-1 P2	E5460 -1 - Cont.-B P2

Example FC Port Connection Table

OSS ID - Port ID	E5460 ID - Port ID
OSS 1 - HBA-1-P1	E5460-1-Cont-A-P1
OSS 1 - HBA-2-P1	E5460-1-Cont-B-P1
OSS 1 - HBA-1-P2	E5460-1-Cont-B-P2
OSS 1 - HBA-2-P2	E5460-1-Cont-A-P2
OSS 2 - HBA-1-P1	E5460-1-Cont-A-P3
OSS 2 - HBA-2-P1	E5460-1-Cont-B-P3
OSS 2 - HBA-1-P2	E5460-1-Cont-B-P4
OSS 2 - HBA-2-P2	E5460-1-Cont-A-P4

Example IB Port Connection Table

OSS ID - Port ID	E5460 ID - Port ID
OSS 1 - HCA-1-P1	E5460-1-Cont-A-IB P1
OSS 1 - HCA-2-P1	E5460-1-Cont-B-IB P1
OSS 1 - HCA-1-P2	E5460-2-Cont-A-IB P1
OSS 1 - HCA-2-P2	E5460-2-Cont-B-IB P1
OSS 2 - HCA-1-P1	E5460-1-Cont-A-IB P2
OSS 2 - HCA-2-P1	E5460-1-Cont-B-IB P2
OSS 2 - HCA-1-P2	E5460-2-Cont-A-IB P2
OSS 2 - HCA-2-P2	E5460-2-Cont-B-IB P2

Example FC Port Connection Table

OSS ID - Port ID	E5460 ID - Port ID
OSS 1 - HBA-1-P1	E5460-1-Cont-A-P1
OSS 1 - HBA-2-P1	E5460-1-Cont-B-P1
OSS 1 - HBA-1-P2	E5460-2-Cont-A-P1
OSS 1 - HBA-2-P2	E5460-2-Cont-B-P1
OSS 2 - HBA-1-P1	E5460-1-Cont-A-P2
OSS 2 - HBA-2-P1	E5460-1-Cont-B-P2
OSS 2 - HBA-1-P2	E5460-2-Cont-A-P2
OSS 2 - HBA-2-P2	E5460-2-Cont-B-P2

Note: The E5424 may replace the E5460 in the configurations shown in Figure 26.

Guidelines

Follow these guidelines when determining the appropriate OSS host configuration for E-Series solutions that use the Lustre file system:

- Carefully plan the primary and secondary OSS connectivity to avoid single points of failure.
- Using the SANtricity ES Host Creation wizard, create a uniquely named host group as part of the process for creating the initial OSS host. Name the group in a manner that facilitates the management of the storage environment over time.

- Map all available LUNs on the storage array to the host group created with the initial OSS host on the array.
- When creating a secondary OSS, select the host group created with the initial host and initiate the Host Creation wizard. The wizard will associate the new host to the existing host group and LUNs on the array.
- Using FC or IB cables, connect the E-Series arrays to the OSSs directly, as shown in Figure 25, to maximize throughput performance, minimize the opportunity for performance bottlenecks, and eliminate single physical points of failure.

  Note: Use the SANtricity ES Linux RDAC multipath driver when there are multiple links from an OSS to an E-Series array.

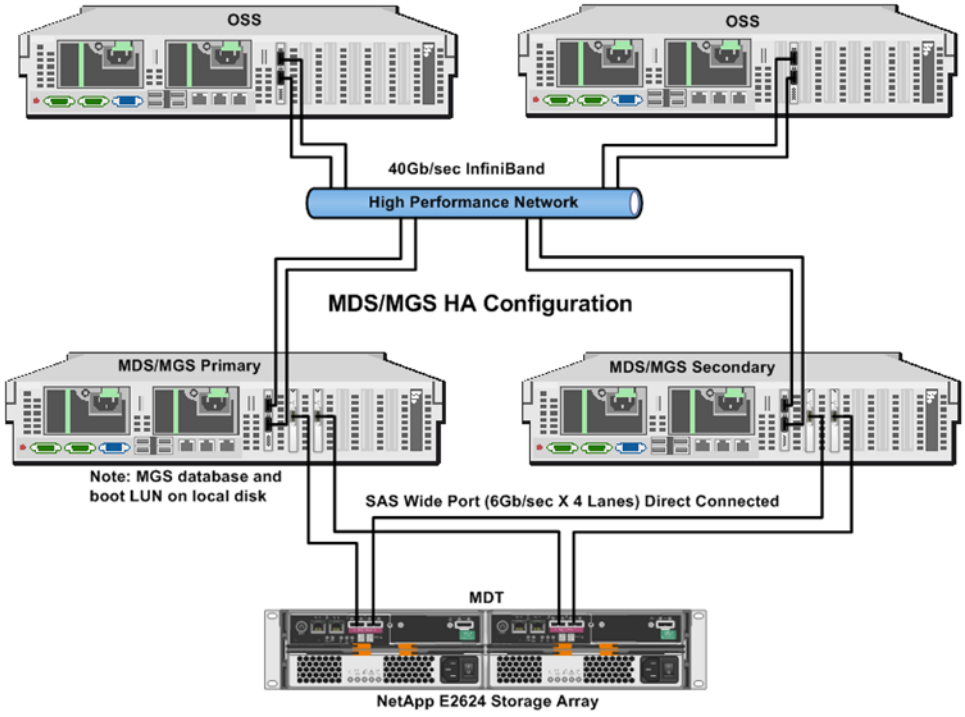
6.2 E-Series MDS and MGS Configuration for Lustre File Systems

Overview

The NetApp HPC Solution for Lustre offers the E-Series E2624 to satisfy the file system MDS storage requirement. The default interface is SAS for MDS-to-MDT access, but the MDSs are IP or IB connected to each other and to all OSSs in the file system for metadata workflows.

Figure 27 provides an example configuration that uses an IB high-speed network for MDS-to-OSS metadata workflows, but the MDS-to-MDT workflows take advantage of the E2600 onboard 4-lane 6Gb/sec SAS interfaces.

Figure 27) Lustre file system redundant MDS HA configuration on E-Series E2624 storage.



To meet the MDS storage access requirement, use SANtricity ES to create two hosts on the E2624 storage array, one per MDS. During the host creation process, when prompted to indicate that the host will be part of a storage partition, create a host group with the primary MDS and enable the MDSs to share access to the MDT through cluster management software. The links from the MDS to the MDT should be configured to avoid single points of failure and to achieve the required throughput levels for the specific Lustre file system implementation.

Note: Each MDS is required to have at least two FC/SAS connections to the MDT for redundancy.

The MGS is typically colocated with the MDS, and the respective targets (MGT and MDT) are also typically colocated on the same storage array. The MDS boot LUN requirement is satisfied by onboard disks in a RAID 1 (1+1) configuration.

Guidelines

Follow these guidelines when determining the appropriate MDS storage configurations for E-Series solutions that use the Lustre file system:

- Install and configure the E-Series array by using standard installation procedures.
- Record the host bus adapter World Wide Port Names (WWPNs) for each MDS port connected to the E-Series array for use when provisioning the hosts on the E2624.
- When creating hosts on the E-Series array, create a host group with the first MDS to allow access for HA MDS-to-MDT workflows.
- Map the MDT LUNs to the MDS host group.
- When creating a secondary MDS host on the E2624, select the existing host group created with the primary MDS and initiate the Host Creation wizard. When prompted, associate the MDS secondary host with the MDS host group.
- Map the FC/SAS links from the MDSs to the E-Series array in a manner that avoids single points of failure and so that each OSS has a path to each E-Series controller.
- Determine the number of required FC/SAS links by using the appropriate sizing guidance from NetApp and Whamcloud for the specific file system environment.
- Connect all MDSs and OSSs by using a private LAN or high-performance network (IB) for metadata workflows between servers.
- Colocate the MGS database with the MDT during the formatting and mounting of the Lustre file system.
- Store the MDS boot LUN on local disks in a RAID 1 (1+1) configuration.

Refer to the Interoperability Matrix Tool (IMT) on the NetApp Support site to validate that the exact product and feature versions described in this document are supported for your specific environment. The NetApp IMT defines the product components and versions that can be used to construct configurations that are supported by NetApp. Specific results depend on each customer's installation in accordance with published specifications.

NetApp provides no representations or warranties regarding the accuracy, reliability, or serviceability of any information or recommendations provided in this publication, or with respect to any results that may be obtained by the use of the information or observance of any recommendations provided herein. The information in this document is distributed AS IS, and the use of this information or the implementation of any recommendations or techniques herein is a customer's responsibility and depends on the customer's ability to evaluate and integrate them into the customer's operational environment. This document and the information contained herein may be used solely in connection with the NetApp products discussed in this document.

Go further, faster®

www.netapp.com